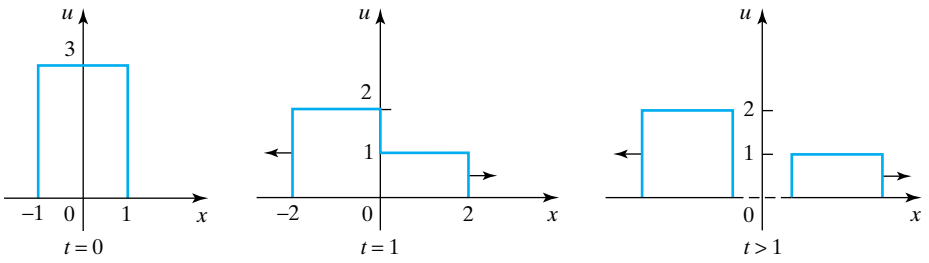


FIGURE 18.15 The resolution of a finite-width initial disturbance into two separate waves that move apart.

ceases. We consider the special case of two initial wave profiles of rectangular shape and the same width, but different heights, that at time $t = 0$ are given by

$$f(x) = \begin{cases} 0, & x < -1 \\ 1, & -1 < x < 1 \\ 0, & x > 1 \end{cases} \quad \text{and } g(x) = \begin{cases} 0, & x < -1 \\ 2, & -1 < x < 1 \\ 0, & x > 1. \end{cases}$$

The evolution of this initial disturbance is shown in Fig. 18.15 for the case $c = 1$. Wave interaction continues until the two disturbances have separated, after which the initial disturbance is represented by two distinct traveling waves.

This result can be explained differently if the wave equation is written in either of the two equivalent forms

$$\left(\frac{\partial}{\partial t} - c\frac{\partial}{\partial x}\right)\left(\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x}\right) = 0 \quad \text{or} \quad \left(\frac{\partial}{\partial t} + c\frac{\partial}{\partial x}\right)\left(\frac{\partial u}{\partial t} - c\frac{\partial u}{\partial x}\right) = 0. \quad (124)$$

An examination of the first of these representations shows that a solution of the *first order* PDE obtained by equating to zero the second group of bracketed terms, namely $u_t + cu_x = 0$, is also a solution of the wave equation. The first order PDE describes a traveling wave of constant shape that propagates to the right at a constant speed c. This special solution is a **degenerate solution** of the wave equation, because it is a solution of a *first order* PDE that is also a *special* solution of a *second order* PDE. Furthermore, unlike the general solution of the wave equation, it is a wave that only moves in *one* direction. When interpreted in terms of the initial conditions used in Fig. 18.15, this degenerate solution is seen to describe the initial wave profile $f(x)$ that after all interaction has ceased becomes the part of the solution of the wave equation that moves to the right.

degenerate solutions

A corresponding argument applied to the other form of the wave equation when the second bracketed term is set equal to zero, so that $u_t + cu_x = 0$, describes a similar degenerate solution that this time moves to the left.

Summary

The wave equation was shown to have a general solution that can be interpreted as the sum of two independent waves moving with the same speed, but in opposite directions. The nature of the solution was used to explain how in wave propagation involving an initial wave profile with discontinuities, the discontinuities propagate along the characteristic curves of the wave equation. A factorization of the wave equation operator was then used to show how special degenerate solutions can arise.

EXERCISES 18.8

In Exercises 1 through 4, the functions $f(x)$ and $g(x)$ refer to the functions in the general solution of the wave equation given in (123). Taking $c = 1$, plot the form of the solution $u(x, t)$ at two different stages during the interaction of the waves, and plot the form of the solution after the waves have separated and all interaction has ceased.

1. $f(x) = \begin{cases} 0, & x < -1 \\ 1 + x, & -1 < x < 1 \\ 0, & x > 1 \end{cases}$,

$g(x) = \begin{cases} 0, & x < -1 \\ 1, & -1 < x < 1 \\ 0, & x > 1. \end{cases}$

2. $f(x) = \begin{cases} 0, & x < -\pi/2 \\ \cos x, & -\pi/2 < x < \pi/2 \\ 0, & x > \pi/2 \end{cases}$,

$g(x) = \begin{cases} 0, & x < -\pi/2 \\ 1 + \dfrac{2}{\pi}x, & -\pi/2 < x < \pi/2 \\ 0, & x > \pi/2. \end{cases}$

3.* $f(x) = \begin{cases} 0, & x < -1 \\ 1 - x^2, & -1 < x < 1 \\ 0, & x > 1 \end{cases}$,

$g(x) = \begin{cases} 0, & x < -1 \\ 1, & -1 < x < 1 \\ 0, & x > 1. \end{cases}$

4.* $f(x) = \begin{cases} 0, & x < -1 \\ 2, & -1 < x < 1 \\ 0, & x > 1 \end{cases}$,

$g(x) = \begin{cases} 0, & x < -1 \\ 1 - x^2, & -1 < x < 1 \\ 0, & x > 1. \end{cases}$

18.9 The D'Alembert Solution of the Wave Equation and Applications

We now derive the promised representation of the solution of the one-dimensional wave equation in terms of its Cauchy conditions that shows explicitly the way in which each of the initial conditions influences the solution. This form of solution is called the **D'Alembert solution**, and the starting point for its derivation is the one-dimensional wave equation for the unknown function $u(x, t)$ where x is a space variable and t is the time.

Let us consider the initial value problem for the homogeneous one-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \quad (c = \text{constant}), \tag{125}$$

subject to the Cauchy conditions

$$u(x, 0) = h(x) \quad \text{and} \quad u_t(x, 0) = k(x), \tag{126}$$

where h and k are suitably differentiable functions defined on the initial line $-\infty < x < \infty$.

It is known from (123) that the general solution of (125) is

$$u(x, t) = f(x - ct) + g(x + ct), \tag{127}$$

where f and g are arbitrary functions of their arguments. Our task will be to find the functions f and g so the solution of the wave equation satisfies the Cauchy conditions in (126). One equation relating f and g follows immediately by setting $t = 0$ in (127) and using the first condition in (126), which gives

$$f(x) + g(x) = h(x). \tag{128}$$

To find another equation we differentiate (127) once partially with respect to t, set $t = 0$, and use the second condition in (126), when we obtain

$$-cf'(x) + cg'(x) = k(x). \tag{129}$$

Integration of (129) from an arbitrary fixed point a on the initial line to a general point x gives

$$-f(x) + g(x) = \frac{1}{c}\int_a^x k(\sigma)d\sigma + g(a) - f(a). \tag{130}$$

Eliminating first $f(x)$ and then $g(x)$ between (128) and (130) gives

$$f(x) = \frac{1}{2}h(x) - \frac{1}{2c}\int_a^x k(\sigma)d\sigma - \frac{1}{2}(g(a) - f(a))$$

and

$$g(x) = \frac{1}{2}h(x) + \frac{1}{2c}\int_a^x k(\sigma)d\sigma + \frac{1}{2}(g(a) - f(a)).$$

If in the expression for $f(x)$ we now replace x by $x - ct$, and in the expression for $g(x)$ we replace x by $x + ct$ and add the results, it follows from (127) that the solution $u(x, t)$ becomes

$$u(x,t) = \frac{1}{2}\left\{h(x-ct) + h(x+ct) - \frac{1}{c}\int_a^{x-ct} k(\sigma)d\sigma + \frac{1}{c}\int_a^{x+ct} k(\sigma)d\sigma\right\}.$$

the D'Alembert solution of the wave equation

Reversing the limits on the first integral, and compensating by changing its sign, allows the two integrals to be combined to give the **D'Alembert solution** of the wave equation:

$$u(x,t) = \frac{h(x-ct) + h(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} k(\sigma)d\sigma. \tag{131}$$

The structure of this solution gives important information about the way the Cauchy conditions enter into the solution of the initial value problem. The implications of (131) can best be understood by interpreting the D'Alembert solution in terms of Fig. 18.16. Consider a representative point P located at (x_0, t_0) in the upper half of the (x, t)-plane, and trace back to the initial line the two characteristics that pass through P with slopes $\pm c$ until they meet the line at points A at $x_0 - ct_0$ and B at $x_0 + ct_0$.

The D'Alembert solution in (131) then shows that the solution at P only depends on the Cauchy conditions over the interval AB on the initial line. Specifically, the solution $u(x_0, t_0)$ only depends on the function $h(x)$ through the two values $h(x_0 - ct_0)$ and $h(x_0 + ct_0)$ at the *ends* of the interval AB, and on $k(x)$ through its integral over the same interval. Because of this, the interval $x_0 - ct_0 \le x \le x_0 + ct_0$ on the initial line is called the **domain of dependence** of the solution at the point (x_0, t_0), and points inside the triangle ABP are said to belong to the **domain of**

domain of dependence and determinacy

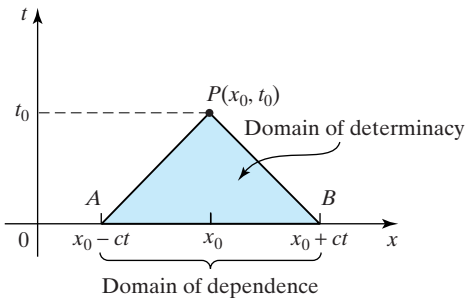


FIGURE 18.16 Domain of dependence and the D'Alembert solution.

determinacy of the interval, because the solution at every point inside this triangle is completely determined by the Cauchy conditions on this interval.

The D'Alembert solution also shows the suitability of Cauchy conditions for the wave equation because they lead to a solution.

The solution is unique, because from the linearity of the wave equation, if two *different* solutions $u(x, t)$ and $v(x, t)$ exist, both satisfying the *same* Cauchy conditions, then the difference between the two solutions $w(x, t) = u(x, t) - v(x, t)$ must also be a solution. The Cauchy conditions for w are $w(x, 0) = 0$ and $w_t(x, 0) = 0$, corresponding to $h(x) \equiv 0$ and $k(x) \equiv 0$, so we conclude from the D'Alembert solution that $w \equiv 0$, and hence that $u \equiv v$.

showing the stability of the solution of a Cauchy problem for the wave equation

We can also use the D'Alembert solution to show the stability of the solution of the wave equation subject to Cauchy conditions, in the sense that a small change in the Cauchy conditions only produces a correspondingly small change in the solution. To show this, let us suppose that $u_1(x, t)$ and $u_2(x, t)$ are two *different* solutions of the wave equation that correspond to the respective *different* Cauchy conditions

$$u_1(x, 0) = h_1(x), \quad u_{1t}(x, 0) = k_1(x), \qquad u_2(x, 0) = h_2(x), \quad \text{and}$$
$$u_{2t}(x, 0) = k_2(x).$$

Now let these two sets of Cauchy conditions be close together in the sense that

$$|h_1(x) - h_2(x)| < \varepsilon_1 \quad \text{and} \quad |k_1(x) - k_2(x)| < \varepsilon_2,$$

where $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ are two arbitrarily small numbers. Applying the elementary integral inequality $|\int_a^b p(x)dx| \le \int_a^b |p(x)|dx$ to this last result gives

$$|u_1(x, t) - u_2(x, t)| < \frac{1}{2}|h_1(x - ct) - h_2(x - ct)| + \frac{1}{2}|h_1(x + ct) - h_2(x + ct)| + \frac{1}{2c}\int_{x-ct}^{x+ct} |k_1(\sigma) - k_2(\sigma)|d\sigma,$$

so as $|k_1(x) - k_2(x)| < \varepsilon_2$ this last result becomes

$$|u_1(x, t) - u_2(x, t)| < \frac{1}{2}\varepsilon_1 + \frac{1}{2}\varepsilon_1 + \frac{\varepsilon_2}{2c}\int_{x-ct}^{x+ct} d\sigma.$$

Finally, after evaluating the integral, we arrive at the result

$$|u_1(x, t) - u_2(x, t)| < \varepsilon_1 + \varepsilon_2 t.$$

This shows that for any time $\tau \leq t$, and arbitrary fixed t, when the two sets of Cauchy data are close together, the corresponding solutions of the wave equation will also be close together, confirming the stability of the solution. The existence of a unique stable solution of the wave equation subject to Cauchy conditions has established that the problem is properly posed.

JEAN-LE-ROND D'ALEMBERT (1717–1783)
A French mathematician born in Paris who was abandoned as a baby near the church of Saint Jean-le-Ronde where he was found by a gendarme who had him christened with the name of the church where he was found. Later, for an unknown reason, he added the name D'Alembert. He was brought up by the wife of a poor glazier, and when he showed early brilliance, his education in law was paid for by his natural father, but his fascination with mathematics was such that he soon abandoned law and devoted himself to the study of mathematics. At the age of 24 he was admitted to the French Academy, and in 1743 he published his great work on mechanics based on what is now known as D'Alembert's principle. He made important contributions to the study of fluid flow, to the study of waves on vibrating strings and elsewhere, and in 1754 made the important suggestion, not to be acted upon until much later, that the then theory of limits needed to be placed on a sound basis. His last years were spent working on the great French encyclopedia.

the solution of the nonhomogeneous wave equation

For reference purposes we state without proof (see, for example, reference [7.20]) that a modification of the preceding argument shows the solution of the **nonhomogeneous wave equation**

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} + f(x,t) \tag{132}$$

is given by

$$u(x,t) = \frac{h(x-ct)+h(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} k(\sigma)d\sigma + \frac{1}{2c}\int_0^t \int_{x-c(t-\tau)}^{x+c(t-\tau)} f(\sigma,\tau)d\sigma\, d\tau. \tag{133}$$

An important and useful result can be derived directly from the general solution of the wave equation in (123), and the fact that its characteristics are $x - ct =$ constant and $x + ct =$ constant. Consider Fig. 18.17, where the four points A at (x_A, t_A), B at (x_B, t_B), C at (x_C, t_C), and D at (x_D, t_D) lie at the corners of a parallelogram, the sides of which are characteristics.

Using the equations of the characteristics, the coordinates of the points A, B, C, and D are seen to be related by

$$\begin{aligned} x_B - ct_B &= x_C - ct_C, \quad & x_A - ct_A &= x_D - ct_D \\ x_A + ct_A &= x_B + ct_B, \quad & x_D + ct_D &= x_C + ct_C. \end{aligned} \tag{134}$$

a useful functional relationship connecting solutions at the corners of a parallelogram formed by characteristic lines

The sums $u(A) + u(C)$ and $u(B) + u(D)$ of the solutions at A, B, C, and D can be written

$$u(A) + u(C) = f(x_A - ct_A) + g(x_A + ct_A) + f(x_C - ct_C) + g(x_C + ct_C)$$

and

$$u(B) + u(D) = f(x_B - ct_B) + g(x_B + ct_B) + f(x_D - ct_D) + g(x_D + ct_D).$$

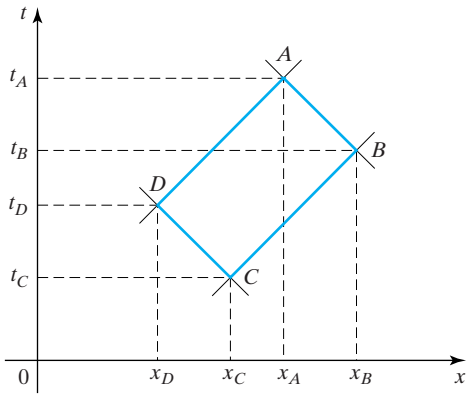


FIGURE 18.17 A parallelogram with sides that coincide with characteristics.

Using the results in (134), we see that these two results are equal, so we have proved that

$$u(A) + u(C) = u(B) + u(D). \tag{135}$$

This result can be used in various ways, one of which is in conjunction with the D'Alembert solution to solve an initial boundary value problem for the wave equation. Let us now find the solution of the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{136}$$

in the quarter-plane $x \geq 0$, $t > 0$ shown in Fig. 18.18, where the solution $u(x, t)$ is required to satisfy the Cauchy conditions $u(x, 0) = h(x)$ and $u_t(x, 0) = k(x)$ on the positive x-axis $x \geq 0$, and the boundary condition $u(0, t) = U(t)$ on the line $x = 0$.

solving an initial boundary value problem

The D'Alembert solution (131) gives the solution in the lower triangular region in Fig. 18.18, but not in the upper triangular region. To find the solution in the upper triangular region we will make use of the D'Alembert solution and result (135).

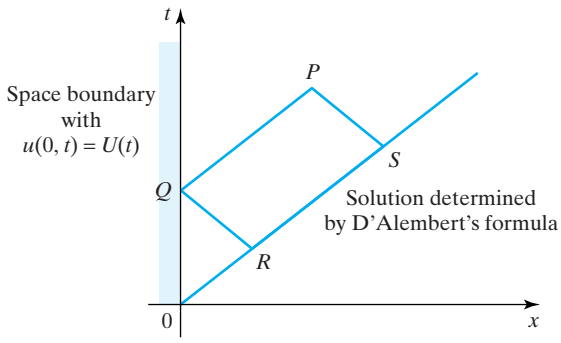


FIGURE 18.18 An initial boundary value problem.

Let P be any point in the upper triangular region, and draw the two characteristics of the wave equation with slopes c and $-c$ that pass through it. Let Q be the point where the characteristic with slope c meets the boundary $x = 0$, and S be the point where the characteristic with slope $-c$ meets the upper boundary of the lower triangular region. Let R be the point where the characteristic through Q with slope $-c$ meets the upper boundary of the lower triangular region. Then, as the sides of the parallelogram $PQRS$ are characteristics, result (135) can be used to relate the solutions at P, Q, R, and S.

The solution $u(x, t)$ at any point in the upper triangular region is now known, because from (135)

$$u(P) = u(Q) + u(S) - u(R),$$

and the solutions at $u(R)$ and $u(S)$ are determined by the D'Alembert solution, while the solution at $u(Q)$ is determined by the given boundary condition $u(0, t) = U(t)$.

This method of solution of an initial boundary value problem in the first quadrant of the (x, t)-plane can be extended to include the case of a semi-infinite strip $a \leq x \leq b, t > 0$ in a straightforward manner, though the details are left as an exercise.

A special case of an initial boundary value problem can be solved by means of the D'Alembert solution without appeal to result (135). To see how this can be done we consider the pure initial value problem for the wave equation

$$u(x, 0) = h(x) \quad \text{and} \quad u_t(x, 0) = k(x), \tag{137}$$

where h and k are bounded odd functions, so that $h(-x) = -h(x)$ and $k(-x) = -k(x)$. Notice that as h and k are odd functions, this implies that $h(0) = k(0) = 0$. The D'Alembert solution applies for all x and $t > 0$, so

$$u(x, t) = \frac{h(x - ct) + h(x + ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} k(\sigma) d\sigma, \tag{138}$$

but as $h(0) = k(0) = 0$, (138) shows that $u(0, t) = 0$.

When in (131) the sign of x is reversed the result becomes

$$u(-x, t) = \frac{h(-x - ct) + h(-x + ct)}{2} + \frac{1}{2c} \int_{-x-ct}^{-x+ct} k(\sigma) d\sigma. \tag{139}$$

However, as h is an odd function, $h(-x - ct) = -h(x + ct)$ and $h(-x + ct) = -h(x - ct)$, so the change of variable $s = -\sigma$ coupled with the fact that k is also an odd function shows that

$$\frac{1}{2c} \int_{-x-ct}^{-x+ct} k(\sigma) d\sigma = -\frac{1}{2c} \int_{x-ct}^{x+ct} k(s) ds.$$

Using these results in (139) and comparing the result with (138) shows that

$$u(-x, t) = -u(x, t). \tag{140}$$

The implication of this result is that if in the D'Alembert solution the Cauchy conditions imposed on the initial line $t = 0$ are such that h and k are *odd* functions,

then if attention is restricted to the *first* quadrant $x \geq 0$, $t > 0$, the D'Alembert solution of this initial value problem solves the initial boundary value problem in which

$$u(x,0) = h(x), \quad u_t(x,0) = k(x) \quad \text{and} \quad u(0,t) = 0, \text{ for } x > 0. \tag{141}$$

reflecting boundary

A useful physical interpretation of this result can be obtained by considering the boundary $x = 0$ to be a **reflecting boundary**, with the property that when a wave moving to the left encounters the boundary it is reflected back in the positive x-direction with a change of sign.

A corresponding result can be derived by assuming h and k to be *even* functions, for then a similar argument shows that the boundary condition imposed on $x = 0$ is the condition

$$u_x(0,t) = 0, \tag{142}$$

but this time when a reflection occurs at the boundary $x = 0$, a wave moving to the left is reflected back in the positive x-direction *without* a change of sign. The details of the proof of this result are left as an exercise.

One-dimensional wave propagation governed by the wave equation is discussed in some detail in references [7.3], [7.10], [7.11], and [7.17] to [7.20].

EXERCISES 18.9

1. Show by differentiation that if f and g are twice differentiable functions of their arguments, $u(x,t) = f(x-ct) + g(x+ct)$ is a solution of the wave equation $u_{tt} = c^2 u_{xx}$.

2. For what value of c is

$$u(x,t) = \frac{1}{2}(x - 4t + 1)e^{-(x-4t)} + \frac{1}{2}(x + 4t - 1)e^{-(x+4t)}$$

a solution of the wave equation $u_{tt} = c^2 u_{xx}$? Find the Cauchy conditions that, when applied to this wave equation, give rise to this solution.

In Exercises 3 through 6 use the D'Alembert solution to solve the given Cauchy problem for the wave equation $u_{tt} = c^2 u_{xx}$.

3. $u(x,0) = \sin x$, $u_t(x,0) = 1/(1+x^2)$.

4. $u(x,0) = 1$, $u_t(x,0) = \cos x$.

5. $u(x,0) = \tanh x$, $u_t(x,0) = \operatorname{sech}^2 x$.

6. $u(x,0) = e^x$, $u_t(x,0) = e^{-x}$.

7. Suggest how the D'Alembert solution and result (135) can be used to solve the initial boundary value problem for the wave equation $u_{tt} = c^2 u_{xx}$ in the semi-infinite strip $a \leq x \leq b$, $t > 0$ when $u(x,0) = h(x)$ with $h(a) = h(b) = 0$, $u_t(x,0) = k(x)$ and $u(a,t) = u(b,0) = 0$. Does this method provide a practical way of solving this initial boundary value problem?

8. By using the form of argument that led to the notion of a reflecting boundary, show that by taking $h(x)$ and $k(x)$ to be *even* functions, the solution given by the D'Alembert formula in the first quadrant solves the initial boundary value problem in that quadrant when

$$u(x,0) = f(x), \quad u_t(x,0) = g(x) \text{ for } x \leq 0,$$

and u satisfies the boundary condition $u_x(0,t) = 0$.

9. Suggest how the D'Alembert solution may be used together with a reflecting boundary to solve the initial boundary value problem in the semi-infinite strip $-a \leq x \leq a$, $t > 0$, subject to the initial and boundary conditions

$$u(x,0) = f(x), u_t(x,0) = g(x) \quad \text{and}$$
$$u(-a,t) = u(a,t) = 0.$$

10. Repeat Exercise 9 with the same initial conditions but with the boundary conditions changed to

$$u(-a,t) = 0 \quad \text{and} \quad u_x(a,t) = 0.$$

11. Write down the D'Alembert solution for the wave equation $u_{tt} = c^2 u_{xx}$ given that the Cauchy conditions are $u(x,0) = f(x)$ and $u_t(x,t) = 0$. Sketch the solution at the times $t = 0$, $1/(2c)$, $1/c$, and $3/(2c)$ using the foregoing initial conditions with

$$f(x) = \begin{cases} 0, & x < -1 \\ -1 - x, & -1 \leq x < 0 \\ 1 - x, & 0 \leq x < 1 \\ 0, & x \geq 1. \end{cases}$$

12. Repeat Exercise 11, but with

$$f(x) = \begin{cases} 0, & x < -1 \\ 1 + x, & -1 \le x < 0 \\ 1 - x, & 0 \le x < 1 \\ 0, & x \ge 1. \end{cases}$$

13. Write down the D'Alembert solution at the time $t = \frac{1}{4}$ for the wave equation $u_{tt} = u_{xx}$, given that the Cauchy conditions are $u(x, 0) = 0$ and $u_t(x, 0) = g(x)$, where

$$g(x) = \begin{cases} 0, & x < -1 \\ 1 - x^2, & -1 \le x \le 1 \\ 0, & x > 1. \end{cases}$$

14. Repeat Exercise 13 with the same Cauchy conditions, but at time $t = \frac{1}{2}$.

18.10 Separation of Variables

The method of solution described in this section applies to homogeneous second and higher order constant coefficient linear PDEs defined in regions D whose spatial boundaries coincide with constant values of the coordinate variables involved. For example, D may be a rectangle with sides parallel to the x-, and y-coordinate axes, a semi-infinite strip parallel to the x-axis, the wedge $r > 0, 0 \le \theta \le \frac{\pi}{4}$ in cylindrical polar coordinates, or the exterior of a sphere of radius R, where it is natural to use spherical polar coordinates with their origin located at the center of the sphere. The success of the method of separation of variables rests on the following results:

1. If u_1 and u_2 are two linearly independent solutions of a homogeneous linear PDE of first or higher order, then the **linear superposition** of the two solutions to give $u = c_1u_1 + c_2u_2$ is also a solution of the PDE, where c_1 and c_2 are arbitrary constants.
2. Under conditions that are satisfied in all ordinary applications, Property 1 extends to the fact that if $u_1, u_2, \ldots,$ is an infinite sequence of linearly independent solutions of a homogeneous linear PDE of second or higher order, then the **linear superposition** of an infinite number of the solutions to give $u = c_1u_1 + c_2u_2 + \cdots$ is also a solution of the PDE, where $c_1, c_2, \ldots$ are arbitrary constants.
3. The orthogonality properties of the eigenfunctions associated with the PDE, special cases of which were developed in Chapter 8, can be used to determine the coefficients $c_1, c_2, \ldots$ in the linear superposition $u = c_1u_1 + c_2u_2 + \cdots$ to make it satisfy the boundary conditions imposed on the PDE, and so become the solution of the boundary value problem.

To illustrate the method we will solve some typical boundary value problems for each of the three fundamental types of second order linear PDE.

Vibrations of a Clamped String

It was shown in Section 18.5 that if a uniform stretched string vibrates in a fixed plane containing its equilibrium position, and the transverse displacement u of the string in this plane remains small, then u must be a solution of the one-dimensional wave equation. If the equilibrium position of the string is taken to coincide with the x-axis and t is the time, the transverse displacement of the string $u(x, t)$ will satisfy the hyperbolic PDE

$$(1/c^2)u_{tt} = u_{xx},$$

where the propagation speed $c = \sqrt{T/\rho}$, with T the tension in the string and ρ the line density of the string.

Let a string of finite length L be clamped rigidly at each end, and choose the origin of the x-axis to coincide with the left end of the string, so its right end will be at the point $x = L$. The *boundary conditions* for the problem then become

$$u(0, t) = u(L, t) = 0, \quad t \geq 0,$$

because these conditions ensure that the ends of the string remain motionless for all time. The Cauchy conditions

$$u(x, 0) = g(x) \quad \text{and} \quad u_t(x, 0) = h(x)$$

determine how the vibration starts at time $t = 0$, with the initial transverse displacement of the string defined by $g(x)$ and its initial transverse speed by $h(x)$. In general the functions g and h are arbitrary, apart from the fact that as the ends of the string are clamped they must be such that $g(0) = g(L) = 0$ and $h(0) = h(L) = 0$.

EXAMPLE 18.11

Consider the vibrations of a stretched string of length L that is clamped at each end and starts from rest with the initial shape $u(x, 0) = kx(L - x)$. Here $k > 0$ is a positive constant chosen such that the maximum transverse displacement is small, in agreement with the approximations made when deriving the wave equation. As the string starts from rest, the Cauchy conditions to be imposed on the wave equation in (143) are

$$u(x, 0) = kx(L - x) \quad \text{and} \quad u_t(x, 0) \equiv 0.$$

the method of separation of variables

The approach to be adopted involves seeking elementary solutions of the wave equation of the form $u(x, t) = X(x)T(t)$, and then using the linearity of the PDE to express the required solution, subject to the boundary and Cauchy conditions, as a linear combination of these elementary solutions. The name **separation of variables** comes from the way the independent variables are *separated* in each elementary solution. In this case the separation involves the product of a function $X(x)$ only of x and a function $T(t)$ only of t.

Partial differentiation of $u(x, t) = X(x)T(t)$ with respect to x only acts on the function $X(x)$, and partial differentiation with respect to t only acts on $T(t)$, so $u_{xx} = X''(x)T(t)$ and $u_{tt} = X(x)T''(t)$, where primes indicate differentiation of the associated function with respect to the appropriate single independent variable.

Substituting these results into the wave equation and dividing by $X(x)T(t)$ gives

$$\frac{1}{c^2}\frac{T''}{T} = \frac{X''}{X}.$$

Inspection of this result shows that the expression on the left is independent of x and so is only a function of t, while the expression on the right is independent of t and so is only a function of x. As x and t are independent variables, the only way a function of t can equal a function of x is if they are each equal to some constant p, so that

$$\frac{1}{c^2}\frac{T''}{T} = \frac{X''}{X} = p,$$

where p is a constant. So T and X must be solutions of the two ordinary differential equations

$$T'' = pc^2T \quad \text{and} \quad X'' = pX.$$

separation constant

The constant p is called a **separation constant**, and before we proceed further it is necessary to determine its *sign*.

Examination of the first equation for $T(t)$ shows that the *time variation* is determined by $T'' = pc^2T$, where $c^2 > 0$, so this equation can only describe *oscillatory behavior* with respect to the time if $p < 0$. Setting $p = -\lambda^2$, with λ a positive real constant, we see that the time variation of the solution is determined by

$$T'' + c^2\lambda^2 T = 0.$$

Our next task will be to find the permissible values of λ, and to do this we must consider the x-variation of the solution that is described by the Sturm–Liouville equation

a Sturm–Liouville problem

$$X'' + \lambda^2 X = 0.$$

The function $X(x)$ determined by this equation must satisfy the boundary conditions on $u(x, t)$ that require $u(0, t) = u(L, t) = 0$. However, as $u(x, t) = X(x)T(t)$ and x and t are independent variables, these boundary conditions on $u(x, t)$ can only hold for all t if $X(0) = X(L) = 0$. This requires that we choose λ so X satisfies the two-point boundary value problem

$$X'' + \lambda^2 X = 0, \text{ with } X(0) = X(L) = 0.$$

This has the general solution

$$X(x) = \widetilde{A}\cos\lambda x + \widetilde{B}\sin\lambda x,$$

where $\widetilde{A}$ and $\widetilde{B}$ are arbitrary constants. Imposing the two-point boundary conditions $X(0) = X(L) = 0$, we have

$$\begin{aligned} &(\text{condition } X(0) = 0) \qquad 0 = \widetilde{A}, \\ &(\text{condition } X(L) = 0) \qquad 0 = \widetilde{B}\sin\lambda L. \end{aligned}$$

The last condition is satisfied if either $\widetilde{B} = 0$, or λL is a zero of the sine function. The condition $\widetilde{B} = 0$ is unacceptable because it makes $X(x)$ identically zero, in which case $u(x, t)$ will also vanish identically, so there can be no vibration of the string. The only alternative is to make λL a zero of the sine function by setting $\lambda L = n\pi$ for $n = 0, 1, 2, \ldots$, where the case $n = 0$ must be omitted because it corresponds to $u(x, t) \equiv 0$. The permissible values of λ, called the **eigenvalues** of the differential equation for $X(x)$, are

$$\lambda_n = \frac{n\pi}{L}, \quad n = 1, 2, \ldots.$$

eigenvalues and eigenfunctions of the Sturm–Liouville problem

The x variation is now seen to be given by

$$X_n(x) = \widetilde{B}\sin\frac{n\pi x}{L}, \quad n = 1, 2, \ldots,$$

where the functions $X_n(x)$ are called the **eigenfunctions** of the differential equation for $X(x)$, and as the equation for X is homogeneous, the value of the constant $\widetilde{B}$ is unimportant.

Once we have determined the permissible values of the eigenvalues λ, the time variation follows by integrating the equation $T'' + c^2\lambda^2 T = 0$, when we find that

$$T_n(t) = \widetilde{C}\cos\frac{nc\pi t}{L} + \widetilde{D}\sin\frac{nc\pi t}{L}, \quad n = 1, 2, \ldots,$$

where the constants $\widetilde{C}$ and $\widetilde{D}$ still remain to be determined. If we substitute for the functions $X_n(x)$ and $T_n(t)$, the permissible elementary solutions become $u_n(x,t) = X_n(x)T_n(t)$ for $n = 1, 2, \ldots$, and these are called the **eigensolutions** of the wave equation. As the constants in $u_n(x,t)$ depend on n, if we replace $\widetilde{B}\widetilde{C}$ by C_n and $\widetilde{B}\widetilde{D}$ by D_n, the eigensolutions of the wave equation become

eigensolutions

$$u_n(x,t) = \sin\frac{n\pi x}{L}\left\{C_n\cos\frac{nc\pi t}{L} + D_n\sin\frac{nc\pi t}{L}\right\},$$

with $n = 1, 2, \ldots$.

Each eigensolution is an elementary solution of the wave equation that satisfies the boundary conditions $u(0,t) = u(L,t) = 0$ for $t \geq 0$, but not the Cauchy conditions. As the wave equation is linear, a linear combination of eigensolutions will also satisfy these same boundary conditions, so we now seek a solution of the initial boundary value problem of the form

$$u(x,t) = \sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty}\sin\frac{n\pi x}{L}\left\{C_n\cos\frac{nc\pi t}{L} + D_n\sin\frac{nc\pi t}{L}\right\},$$

where the coefficients C_n and D_n are to be chosen so that $u(x,t)$ satisfies the Cauchy conditions

$$u(x,0) = kx(L-x) \quad \text{and} \quad u_t(x,0) \equiv 0.$$

To find the coefficients C_n and D_n we need to make use of Fourier series. First setting $t = 0$ in the expression for $u(x,t)$ and using the first initial condition gives

$$kx(L-x) = \sum_{n=1}^{\infty} C_n\sin\frac{n\pi x}{L}.$$

Then, assuming that differentiation of the series for $u(x,t)$ with respect to t is permissible, setting $t = 0$ in the result, and using the second initial condition gives

$$0 = \frac{c\pi}{L}\sum_{n=1}^{\infty} nD_n\sin\frac{n\pi x}{L}.$$

The series involving the coefficients C_n and D_n are simply the Fourier sine series expansion of the functions on the left, so it follows immediately that $D_n = 0$ for $n = 1, 2, \ldots$. To find the coefficients C_n we multiply series for $kx(L-x)$ by $\sin m\pi x/L$ and integrate from $x = 0$ to $x = L$, when we obtain

$$\begin{aligned}\int_0^L kx(L-x)\sin\frac{m\pi x}{L}dx &= \int_0^L\sum_{n=1}^{\infty} C_n\sin\frac{n\pi x}{L}\sin\frac{m\pi x}{L}dx \\ &= \sum_{n=1}^{\infty}\int_0^L C_n\sin\frac{n\pi x}{L}\sin\frac{m\pi x}{L}dx,\end{aligned}$$

where the justification for the interchange of the summation and integral signs has been omitted. As the set of functions $\{\sin(m\pi x/L)\}_{m=1}^{\infty}$ is orthogonal on the interval

$0 \le x \le L$, the preceding result reduces to

$$\int_0^L kx(L-x)\sin\frac{m\pi x}{L}dx = C_m\int_0^L \sin^2\frac{m\pi x}{L}dx.$$

After the integrations are performed, this becomes

$$\left(-\frac{2kL^3}{m^3\pi^3}\cos m\pi + \frac{2kL^3}{m^3\pi^3}\right) = \frac{L}{2}C_m \quad \text{for} \quad m = 1, 2, \ldots.$$

Using the result $\cos m\pi = (-1)^m$, we see that the expression on the left vanishes when m is even, so setting $m = 2r$ with $r = 1, 2, \ldots$, we have $C_{2r} = 0$. However, when m is odd the expression on the left no longer vanishes, and setting $m = 2r + 1$ with $r = 0, 1, \ldots$ simplifies the result to

$$\frac{4kL^3}{(2r+1)^3\pi^3} = \frac{L}{2}C_{2r+1}.$$

The coefficients C_r are now all known and are given by

$$C_{2r} = 0 \quad \text{and} \quad C_{2r+1} = \frac{8kL^2}{(2r+1)^3\pi^3} \quad \text{for } r = 0, 1, \ldots.$$

Substituting for the coefficients C_n in the series for $u(x, t)$, and setting the coefficients $D_n = 0$, we arrive at the required solution

$$u(x,t) = \frac{8kL^2}{\pi^3}\sum_{r=0}^{\infty}\frac{1}{(2r+1)^3}\sin\frac{(2r+1)\pi x}{L}\cos\frac{(2r+1)c\pi t}{L},$$

for $0 \le x \le L$ and $t \ge 0$.

The justification for differentiating the functional series $u(x, t)$ term by term with respect to t and for interchanging the summation and integral signs requires arguments involving uniform convergence and so will be omitted.

It is instructive to interpret the eigenfunctions $X_n(x)$ and the eigensolutions $u_n(x, t)$ in physical terms. Inspection of the solution shows that the eigenfunction $X_n(x)$ defines the nth **mode** of the vibration, in the sense that however $X_n(x)$ is scaled, it always specifies the *shape* of the string corresponding to a given value of n. The nth eigensolution $u_n(x, t)$ is seen to be the time modulation of the nth mode. This describes how the nth mode vibrates with time and shows that it experiences a periodic variation of amplitude and a change of sign. The solution is a linear combination of all of the possible modes of vibration, chosen such that when $t = 0$ the shape of the string is $u(x, 0) = kx(L - x)$.

modes of vibration

If the initial shape of the string is changed, but the second Cauchy condition $u_t(x, 0) \equiv 0$ is retained, the new solution will simply be a *different* linear combination of these *same* eigensolutions.

Figure 18.19 shows the initial shape of the string at time $t = 0$, and its shape at three subsequent times where, for convenience, we have set $L = \pi$, $c = 1$ and graphed the approximation to the function $\hat{u} = (\frac{\pi}{8k})u(x, t)$ using only the first 10 terms of the series solution. ■

Vibrations of a Circular Membrane

To illustrate the method of separation of variables when applied to the wave equation in more than one space variable, we will examine the vibrations of a uniform

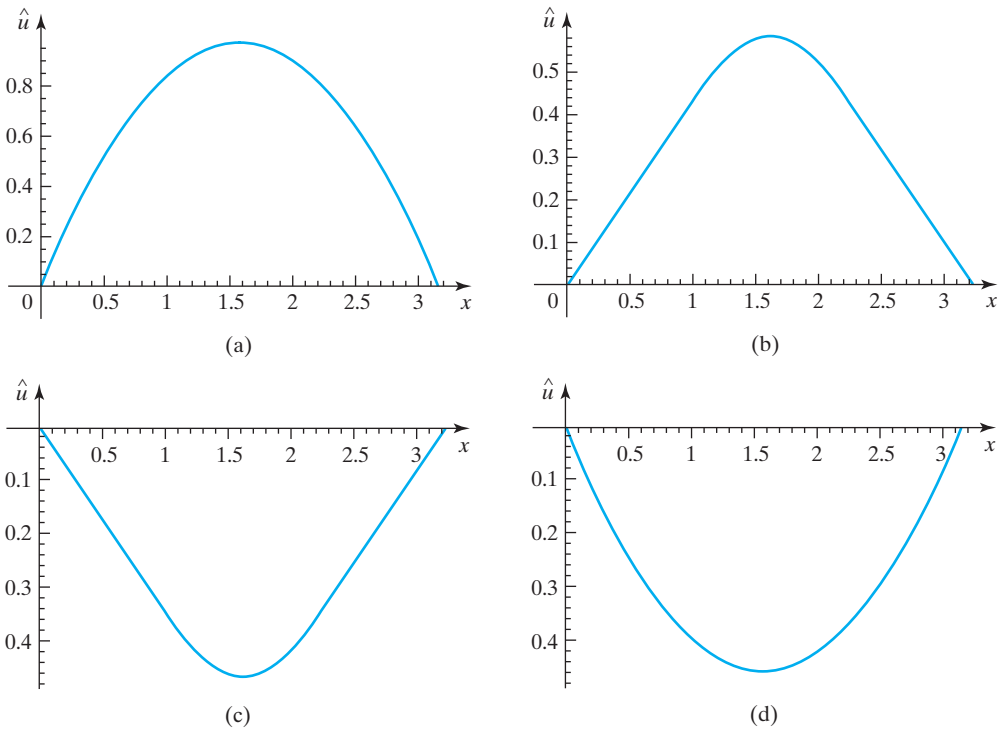


FIGURE 18.19 The shape of the string at different times (a) $t = 0$, (b) $t = 1$, (c) $t = 2$, (d) $t = 3$.

circular membrane of unit radius clamped around its rim. Because of the circular boundary, when we solve this problem the two space variables will be taken to be the cylindrical polar coordinates (r, θ), with their origin at the center of the membrane when in its equilibrium position, and the third independent variable will be the time t. The displacement of the membrane normal to its equilibrium position will be denoted by $u(r, \theta, t)$.

This problem can be considered to be a mathematical description of the vibrations of a circular membrane covering a drum that is subjected to Cauchy conditions at an initial time $t = 0$ that describe the vertical displacement $u(r, \theta, t)$ and the speed $u_t(r, \theta, t)$ of the membrane in a direction normal to its equilibrium position. It will be shown that the response to arbitrary Cauchy conditions is expressible as a sum of eigensolutions in a manner analogous to that of the vibrating string.

EXAMPLE 18.12

a vibration problem involving cylindrical polar coordinates

The geometry of the circular membrane problem suggests that cylindrical polar coordinates should be used. When the wave equation is expressed in terms of cylindrical polar coordinates it becomes

$$u_{tt} = c^2\left(u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}\right) \quad \text{or} \quad u_{tt} = c^2\Delta u,$$

where in cylindrical polar coordinates the Laplacian $\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}$.

The boundary conditions are

$$u(1, \theta, t) = 0 \text{ for } 0 \le \theta \le 2\pi \text{ and } t > 0 \quad \text{(the rim is clamped)}$$

and

$$u(r, \theta, t) \text{ is finite for } 0 \le r \le 1 \text{ and } t > 0 \quad \text{(the displacement is finite)},$$

while the initial, or Cauchy, conditions describing the initial shape of the membrane and its initial speed normal to its equilibrium position are

$$u(r, \theta, 0) = f(r, \theta) \quad \text{and} \quad u_t(r, \theta, 0) = g(r, \theta).$$

It will be simplest if the variables are separated in two stages, so first we separate out the time t by setting $u(r, \theta, t) = H(r, \theta)T(t)$, and then substitute into the differential equation to obtain

$$HT'' = c^2 T \nabla^2 H,$$

where primes denote differentiation with respect to the independent variable t occurring in $T(t)$. Dividing by HT, we have

$$\frac{1}{c^2}\frac{T''}{T} = \frac{\nabla^2 H}{H},$$

but as the expression on the left is a function of the independent time variable t, and the one on the right is a function of the independent space variables r and θ, this can only be true if

$$\frac{1}{c^2}\frac{T''}{T} = \frac{\nabla^2 H}{H} = k,$$

where k is a constant. The time variation is determined by $T'' - c^2 kT = 0$, so for the solution to be periodic in time, as is necessary if it is to describe vibrations, it is necessary that $k < 0$. Accordingly, if we set the separation constant $k = -\lambda^2$, the equations for T and H become

$$T'' + \lambda^2 c^2 T = 0$$

and

$$\Delta^2 H + \lambda^2 H = 0.$$

Helmholtz equation

The partial differential equation for H is called the **Helmholtz equation**, and it plays a fundamental role in studies of the wave equation. To find the permissible values of the eigenvalues λ we must now solve the Helmholtz equation, because the eigenvalues will be determined by the boundary conditions that must be imposed on H.

To this end we set $H(r, \theta) = R(r)\Theta(\theta)$, and after substituting for H in the Helmholtz equation we obtain

$$\Theta\left(R'' + \frac{1}{r}R'\right) + \frac{R}{r^2}\Theta'' + \lambda^2 R\Theta = 0.$$

Dividing this result by $R\Theta$ and rearranging terms gives

$$\frac{r^2}{R}\left(R'' + \frac{1}{r}R'\right) + \lambda^2 r^2 = -\frac{\Theta''}{\Theta}.$$

The expression on the left is only a function of the independent variable r, and the one on the right is only a function of the independent variable θ, so this can only be possible if

$$\frac{r^2}{R}\left(R'' + \frac{1}{r}R'\right) + \lambda^2 r^2 = -\frac{\Theta''}{\Theta} = m,$$

where m is another separation constant. The preceding result can now be decoupled to give the two Sturm–Liouville equations for $R(r)$ and $\Theta(\theta)$

$$r^2 R'' + r R' + (\lambda^2 r^2 - m)R = 0 \quad \text{and} \quad \Theta'' + m\Theta = 0.$$

To solve these equations it is necessary to supply boundary conditions for both R and Θ. As the variables are separable, these conditions follow if we interpret the boundary conditions for $u(r, \theta, t)$ in terms of $H(r, \theta) = R(r)\Theta(\theta)$. The boundary conditions give rise to two conditions, the first of which corresponds to the clamping of the rim that can be expressed by the requirement $R(1) = 0$, which ensures that the rim of the membrane remains fixed at all times. The second condition, which at first sight appears a little strange, is the requirement that $R(r)$ be *finite* for $0 \le r \le 1$. The need for this seemingly obvious requirement will become clear later.

The condition to be imposed on θ follows from the fact that the membrane is circular, so for the solution to have circular symmetry θ must be periodic with period 2π. The equation for Θ can only give rise to solutions that are periodic if $\sqrt{m} > 0$, in which case the solution becomes

$$\Theta(\theta) = \widetilde{A}\cos(\sqrt{m}\theta + \phi),$$

where $\widetilde{A}$ and ϕ are arbitrary constants. This solution will only be periodic with period 2π, as is required by the nature of the problem, if $\sqrt{m} = n$ for $n = 0, 1, \ldots$, so setting $m = n^2$, we see that the angular variation is determined by

$$\Theta(\theta) = \widetilde{A}\cos(n\theta + \phi).$$

The choice of reference line through the origin relative to which the polar angle θ is measured is immaterial, so without loss of generality it will be chosen to make the constant $\phi = 0$, because then the angular variation is determined by

$$\Theta(\theta) = \widetilde{A}\cos(n\theta).$$

If we substitute $m = n^2$ for the separation constant, the radial variation is seen to be governed by *Bessel's equation*

$$r^2 R'' + r R' + (\lambda^2 r^2 - n^2)R = 0 \quad \text{for } 0 < r < 1, n = 0, 1, 2, \ldots.$$

how Bessel's equation and its zeros enter into this solution of the wave equation

The general solution of this form of Bessel's equation (see Sections 8.6 and 8.7) is

$$R(r) = \widetilde{B}J_n(\lambda r) + \widetilde{C}Y_n(\lambda r),$$

and to determine the two arbitrary constants $\widetilde{B}$ and $\widetilde{C}$ we now make use of the two boundary conditions for $R(r)$ that were found earlier. The need for the condition

that $R(r)$ remains finite for $0 \le r \le 1$ will be used first. This boundary condition shows that the term $Y_n(\lambda r)$ must be omitted from the solution $R(r)$ if u is to remain finite when $r = 0$, because $Y_n(x)$ is infinite at the origin. So we must set $\widetilde{C} = 0$, when the radial variation becomes

$$R(r) = \widetilde{B} J_n(\lambda r).$$

The permissible values of λ now follow by using the remaining boundary condition $R(1) = 0$. This condition shows that we must set $J_n(\lambda) = 0$, so λ must be one of the infinite number of nonvanishing zeros of $J_n(x)$. If we denote these by $j_{n,s}$ for $s = 1, 2, \ldots$, the eigenvalues λ must be

$$\lambda = j_{n,s}.$$

A listing of the first few of these zeros is given in Section 8.6.

Combining the foregoing results shows that the eigenfunction determining the (n, s)-mode of vibration is

$$H_{ns}(r, \theta) = \widetilde{A}\widetilde{B} J_n(j_{n,s} r) \cos(n\theta),$$

where the product of the arbitrary constants, itself another arbitrary constant, will depend on n and s.

The time variation follows by integrating $T'' + \lambda^2 c^2 T = 0$, when we find that

$$T(t) = \widetilde{D} \cos(j_{n,s} ct) + \widetilde{E} \sin(j_{n,s} ct),$$

where here also the two arbitrary constants depend on n and s.

Finally, combining results to obtain a general eigensolution gives

$$u_{ns}(r, \theta, t) = J_n(j_{n,s} r) \cos(n\theta)\{P_{ns} \cos(j_{n,s} ct) + Q_{ns} \sin(j_{n,s} ct)\}.$$

Here, because the arbitrary constants depend on n and s, and a product of arbitrary constants is also an arbitrary constant, we have set $P_{ns} = \widetilde{A}\widetilde{B}\widetilde{D}$ and $Q_{ns} = \widetilde{A}\widetilde{B}\widetilde{E}$.

typical vibrational modes and nodal lines

Before we solve the initial value problem, let us first examine the nature of the eigenfunctions $H_{ns}(r, \theta)$ that determine the *shape* of each mode of vibration. As $H_{ns}(r, \theta)$ is modulated by the time variation $T(t)$, the general shape of the (n, s)-mode can be seen by setting the product of arbitrary constants equal to 1 and taking the eigenfunction to be $H_{ns}(r, \theta) = J_n(j_{n,s} r) \cos(n\theta)$. The diagrams in Fig. 18.20 illustrate the first few vibrational modes. The shaded and unshaded areas in the diagrams indicate where displacement occurs in *opposite* directions. The modulation of an eigenfunction by the time variation $T(t)$ simply alters the amplitude of the displacement, and periodically reverses its direction. The lines bordering the shaded and unshaded areas are called **nodal lines**, and these represent lines on the surface of the membrane that are never displaced from their equilibrium position. As n and s increase, so also does the complexity of the pattern of the nodal lines. Figure 18.21a illustrates the membrane displacement in the eigenmode corresponding to $n = 2$ and $s = 1$, and Fig. 18.21b shows the corresponding contour lines.

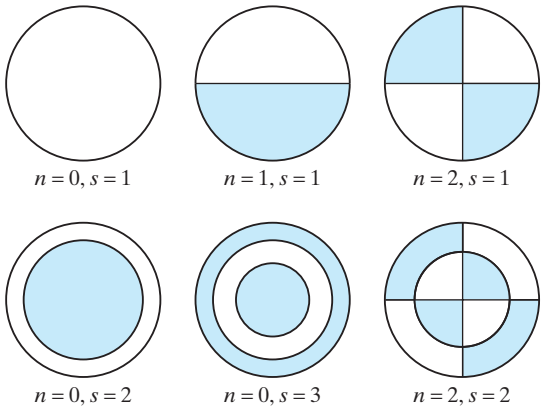


FIGURE 18.20 Some typical vibrational modes.

As with the stretched string, we now express the required solution that satisfies the Cauchy conditions as the linear combination of eigensolutions

$$u(r, \theta, t) = \sum_{n=0,s=1}^{\infty} u_{ns}(r, \theta, t).$$

Substituting for $u_{ns}(r, \theta, t)$ gives

$$u(r, \theta, t) = \sum_{n=0,s=1}^{\infty} J_n(j_{n,s}r)\cos(n\theta)\{P_{ns}\cos(j_{n,s}ct) + Q_{ns}\sin(j_{n,s}ct)\}.$$

To satisfy the Cauchy conditions it is necessary to set $u(r, \theta, 0) = f(r, \theta)$ and $u_t(r, \theta, 0) = g(r, \theta)$, and then to solve for the coefficients P_{ns} and Q_{ns}. To do this we will make use of the orthogonality of the set of cosine functions $\{\cos(n\theta)\}|_{n=0}^{\infty}$ over the interval $0 \le \theta \le 2\pi$ and the orthogonality of the set of Bessel functions

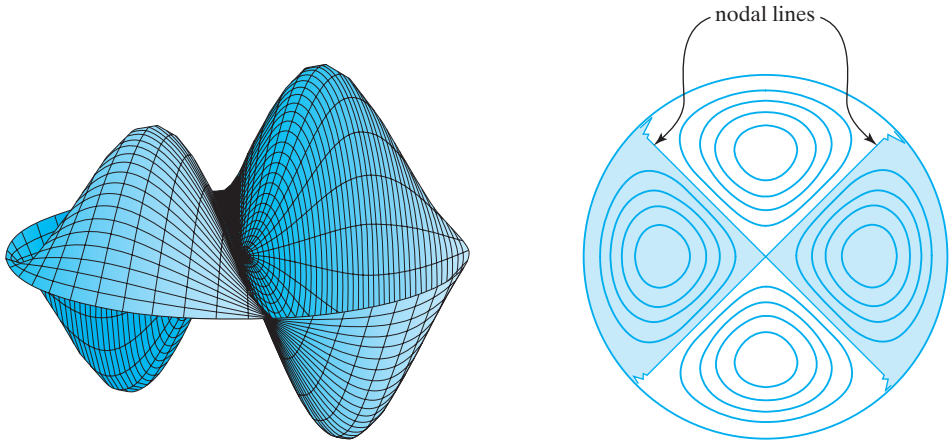


FIGURE 18.21 (a) The membrane displacement and (b) a contour plot for $H_{21}(r, \theta)$.

$\{J_m(j_{m,q}r)\}|_{q=1}^{\infty}$ over the interval $0 \le r \le 1$, and when doing so we will make use of the results of Example 8.25, where it was shown that

$$\int_0^1 r J_m(j_{m,p}r) J_m(j_{m,q}r)dr = \begin{cases} 0, & p \neq q \\ \dfrac{1}{2}[J_{m+1}(j_{m,q})]^2, & p = q. \end{cases}$$

Using the first Cauchy condition, setting $t = 0$, multiplying the result by $rJ_m(j_{m,q}r)\cos(m\theta)$, and integrating with respect to r over the interval $0 \le r \le 1$, and then with respect to θ over the interval $0 \le \theta \le 2\pi$ gives

using the orthogonality of Bessel functions to determine the coefficients

$$\int_0^1 \int_0^{2\pi} r J_m(j_{m,q}r)\cos(m\theta) f(r,\theta)d\theta dr$$
$$= \sum_{n=0,s=1}^{\infty} P_{ns} \int_0^1 \int_0^{2\pi} r J_m(j_{m,q}r) J_n(j_{n,s}r)\cos(m\theta)\cos(n\theta)d\theta dr.$$

The orthogonality properties of the Bessel and cosine functions in the series on the right cause all but the term in P_{mq} to vanish, so that the result reduces to the single term

$$\int_0^1 \int_0^{2\pi} r J_m(j_{m,q}r)\cos(m\theta) f(r,\theta)d\theta dr$$
$$= P_{mq}\left\{\int_0^1 r[J_m(j_{m,q}r)]^2 dr\right\}\left\{\int_0^{2\pi} \cos^2(m\theta)d\theta\right\}.$$

Evaluating the integrals and solving for P_{mq}, we find that

$$P_{0q} = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} r J_0(j_{0,q}r) f(r,\theta)d\theta dr/[J_1(j_{0,q})]^2 \quad \text{for } m = 0, q = 1, 2, \ldots,$$

and

$$P_{mq} = \frac{2}{\pi}\int_0^1 \int_0^{2\pi} r J_m(j_{m,q}r)\cos(m\theta) f(r,\theta)d\theta dr/[J_{m+1}(j_{m,q})]^2 \text{ for } m, q = 1, 2, \ldots.$$

Differentiation of $u(r,\theta,t)$ with respect to t, followed by setting $t = 0$, shows that after setting $u_t(r,\theta,0) = g(r,\theta)$ we obtain

$$g(r,\theta) = \sum_{n=0,s=1}^{\infty} Q_{ns} j_{n,s} J_n(j_{n,s}r)\cos(n\theta).$$

The coefficients Q_{ns} can be found in the same way as the coefficients P_{ns}, and the formulas for them follow from the results for P_{ns} by replacing $f(r,\theta)$ by $g(r,\theta)$.

If the vibrations are circularly symmetric, and so do not depend on θ, the expression $u(r,\theta,0)$ simplifies to $u(r,\theta,0) = h(r)$, say. If, in addition, the vibrations start from rest, so $u_t(r,\theta,0) = 0$, the solution simplifies still further, because then $m = 0$ and only the coefficients P_{0q} are nonvanishing, so that

$$P_{0q} = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} r J_0(j_{0,q}r)h(r)d\theta dr/[J_1(j_{0,q})]^2.$$

After integrating with respect to θ (that introduces a factor 2π), we find that

$$P_{0q} = 2\int_0^1 r J_0(j_{0,q}r)h(r)dr/[J_1(j_{0,q})]^2 \quad \text{for } q = 1, 2, \ldots.$$

In terms of these coefficients the solution then takes the particularly simple form

$$u(r,t) = \sum_{q=1}^{\infty} J_0(j_{0,q}r)\,P_{0q}\cos(j_{0,q}ct) \quad \text{for } 0 \le r \le 1,\ t > 0. \qquad \blacksquare$$

This same method of analysis can be used when the membrane is in the form of an annulus $r_1 \le r \le r_2$, with Dirichlet and/or Neumann conditions imposed on its inner and outer boundaries. In this case the solution is not required at the origin $r = 0$, so the term $Y_n(r)$ must be retained in the solution for $R(r)$, which then becomes $R(r) = \widetilde{B}J_n(\lambda r) + \widetilde{C}Y_n(\lambda r)$. The eigenvalues λ_n follow by applying the appropriate boundary conditions to $R(r)$ at $r = r_1$ and $r = r_2$, but depending on the boundary conditions the determination of the numerical values of the eigenvalues can be difficult, so it is usually necessary to obtain them by numerical methods.

Time Variation of Temperature in a Long Thin Metal Plate or Rod

The following example illustrates how the method of separation of variables can be applied to a time-dependent heat flow problem in a long thin metal plate of width L.

EXAMPLE 18.13

We consider the long thin metal plate of width L in the x-direction illustrated in Fig. 18.22, with negligible thickness in the y-direction and a length in the z-direction that is much greater than L. The edge $x = 0$ is kept at zero temperature and the edge $x = L$ is thermally insulated, so no heat can pass through it. The temperature distribution across the width of the plate will be assumed to be independent of z, so as the thickness in the y-direction is negligible, the temperature distribution will depend only on x and t. The initial temperature distribution across the width of the plate applied at $t = 0$ will be taken to be $u(x, 0) = u_0(1 + x/L)$.

As the temperature distribution across the plate will be the same in any plane $z =$ constant, this situation also models a rod of length L in the plane $z = 0$, along the x-axis, when its faces above and below the plane $z = 0$ are thermally insulated. In each case the temperature distribution $u(x, t)$ will be determined by the one-dimensional heat equation

$$u_t = \kappa^2 u_{xx}.$$

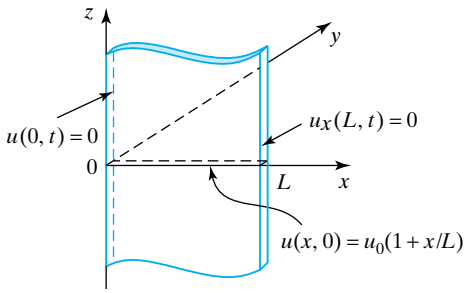


FIGURE 18.22 The plate of width L and the boundary and initial conditions.

Solution The boundary conditions on the plate are

$$u(0,t) = 0 \quad \text{and} \quad u_x(L,t) = 0 \quad \text{for } t > 0,$$

where the first condition says that the left edge of the plate is maintained at zero temperature, and the second says that there is no heat flux across the edge $x = L$. The initial condition to be imposed across the plate is

$$u(x,0) = u_0(1 + x/L).$$

Setting $u(x,t) = X(x)T(t)$, substituting into the heat equation, and dividing by XT gives

$$\frac{T'}{T} = \kappa^2 \frac{X''}{X}.$$

As the expression on the left is only a function of t and the one on the right is only a function of x, this can only be possible if

$$\frac{T'}{T} = \kappa^2 \frac{X''}{X} = k,$$

where k is a separation constant. To determine the sign of k, we appeal to the physical condition that the temperature cannot become infinite with the increase of time, so as $T' = kT$, this can only be possible if $k < 0$, so we set $k = -\lambda$ with $\lambda > 0$. The differential equations governing T and X now become

$$X'' + \frac{\lambda}{\kappa^2} X = 0 \quad \text{and} \quad T' + \lambda T = 0,$$

so the X variation is given by

$$X(x) = \widetilde{A}\cos(\sqrt{\lambda}x/\kappa) + \widetilde{B}\sin(\sqrt{\lambda}x/\kappa).$$

The boundary conditions for X follow from the boundary conditions for the temperature, so as the variables are separable, we require that $X(0)T(t) = 0$ and $X'(L)T(t) = 0$ for $t > 0$. Thus, the boundary conditions on X must be $X(0) = 0$ and $X'(L) = 0$. The equation for $X(x)$ is a Sturm–Liouville problem, so applying these boundary conditions gives

$$\begin{aligned} &\text{(the condition } X(0) = 0) && 0 = \widetilde{A} \\ &\text{(the condition } X'(L) = 0) && 0 = \frac{\sqrt{\lambda}}{\kappa}\widetilde{B}\cos(\sqrt{\lambda}L/\kappa). \end{aligned}$$

If $\widetilde{B} = 0$, the solution vanishes identically, so as this is impossible, the eigenvalues λ must be solutions of

$$\cos(\sqrt{\lambda}L/\kappa) = 0,$$

which are the zeros of the cosine function

$$\frac{\sqrt{\lambda_n}}{\kappa} L = (2n+1)\frac{\pi}{2}, \quad \text{or} \quad \lambda_n = \frac{(2n+1)^2\pi^2\kappa^2}{4L^2} \quad \text{for } n = 0, 1, \ldots.$$

The eigenfunctions are thus

$$X_n(x) = \widetilde{B}\sin(2n+1)\frac{\pi x}{2L} \quad \text{for } n = 0, 1, \ldots,$$

and the time variation of the eigenfunctions follows if we integrate

$$T' + \frac{(2n+1)^2\pi^2\kappa^2}{4L^2} T = 0$$

to obtain

$$T_n(t) = \widetilde{C} \exp\left[-\frac{(2n+1)^2\pi^2\kappa^2 t}{4L^2}\right].$$

If we set $C_n = \widetilde{B}\widetilde{C}$, because both coefficients depend on n, the nth eigensolution becomes

$$u_n(x,t) = C_n \sin(2n+1)\frac{\pi x}{2L} \exp\left[-\frac{(2n+1)^2\pi^2\kappa^2 t}{4L^2}\right], \quad \text{for } n = 0, 1, \ldots.$$

We now seek a solution in the form of the linear combination of eigensolutions

$$u(x,t) = \sum_{n=0}^{\infty} u_n(x,t) = \sum_{n=0}^{\infty} C_n \sin(2n+1)\frac{\pi x}{2L} \exp\left[-\frac{(2n+1)^2\pi^2\kappa^2 t}{4L^2}\right].$$

To determine the coefficients C_n it is necessary to make use of the initial condition $u(x,0) = u_0(1 + x/L)$. Setting $t = 0$ in this expression and using the initial condition gives

$$u_0(1 + x/L) = \sum_{n=0}^{\infty} C_n \sin(2n+1)\frac{\pi x}{2L}.$$

Multiplying this result by $\sin(2m+1)\frac{\pi x}{2L}$, integrating with respect to x over the interval $0 \le x \le L$, and using the orthogonality properties of the set of functions $\left\{\sin(2n+1)\frac{\pi x}{2L}\right\}$ leads to the equation for C_n

$$u_0 \int_0^L (1 + x/L) \sin(2n+1)\frac{\pi x}{2L} dx = C_n \int_0^L \left[\sin(2n+1)\frac{\pi x}{2L}\right]^2 dx.$$

Evaluating the integrals and then solving for C_n, we have

$$C_n = \frac{4u_0}{\pi(2n+1)}\left[1 + (-1)^n \frac{2}{(2n+1)\pi}\right] \quad \text{for } n = 0, 1, \ldots,$$

and the solution now follows if we substitute this expression for C_n into the series solution for $u(x,t)$.

A computer plot of $\hat{u}(x,t)/u_0$, obtained by using the first 50 terms in the series solution with $L = 1$ and $\kappa = 1$, is shown in Fig. 18.23. This confirms, as expected, that the solution decays to zero as t increases. The scale of the plot is too small to show the Gibbs phenomenon near $x = 0, t = 0$ where there is a discontinuity. ■

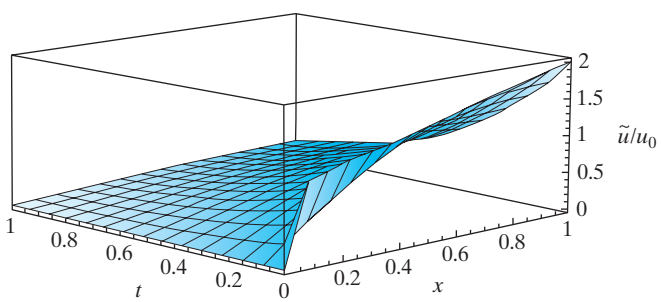


FIGURE 18.23 A plot of $\hat{u}(x,t)/u_0$ for $0 \le x \le 1, t > 0$.

The boundary conditions used so far have been particularly simple, but in physical situations they are often more complicated, and instead of being either a Dirichlet or Neumann condition they may involve a linear combination of both of these conditions. For example, a condition of the form $(\partial u/\partial x + Ku)|_{x=a} = f(t)$ describes how a combination of the temperature and heat flux is required to vary as a function of time t at the boundary $x = a$. The next example involves a boundary condition of this type, and it demonstrates how under such conditions the eigenvalues can become the zeros of a transcendental equation, and so must be found numerically.

EXAMPLE 18.14

Solve the heat equation

$$\frac{\partial u}{\partial t} = \kappa^2 \frac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le L,\ t > 0,$$

subject to the boundary conditions

$$u(0, t) = 0 \quad \text{and} \quad \left(\frac{\partial u}{\partial x} + Ku\right)\bigg|_{x=L} = 0,\ K > 0,$$

and the initial condition

$$u(x, 0) = \sin(\pi x/L).$$

Solution Separating variables by seeking elementary solutions of the form $u(x, t) = X(x)T(t)$, substituting into the heat equation, and dividing by $X(x)T(t)$, we obtain

$$\frac{X''}{X} = \frac{1}{\kappa^2}\frac{T'}{T} = -\lambda^2,$$

where λ^2 is a positive real separation constant. So, as usual, we arrive at the two ordinary differential equations

$$X'' + \lambda^2 X = 0 \quad \text{and} \quad T' + \lambda^2 \kappa^2 T = 0,$$

the first of which is a Sturm–Liouville problem.

The general solution for $X(x)$ is $X(x) = A\cos\lambda x + B\sin\lambda x$. The boundary condition $u(0, t) = 0$ shows that $A = 0$, so $X(x) = B\sin\lambda x$, while the boundary condition $(\partial u/\partial x + Ku)|_{x=L} = 0$ leads to the condition

$$\lambda B\cos\lambda L + KB\sin\lambda L = 0, \quad \text{and so} \quad \tan\lambda L = -\lambda/K.$$

a transcendental equation for the eigenvalues

Setting $\mu = \lambda L$ and $p = KL > 0$, we find that the eigenvalues μ are determined by the zeros of the transcendental equation

$$\tan\mu = -\frac{\mu}{p}.$$

The positive values of μ can be estimated from the points of intersection the graphs of $y = \tan\mu$ and $y = -\mu/p$ for $\mu > 0$. Figure 18.24 shows a typical case when $p = 1$.

Denoting the positive roots (the eigenvalues) of this equation by $\mu_1, \mu_2, \ldots$ and solving the time variation equation $T_n' + \lambda_n^2\kappa^2 T_n = 0$ gives

$$T_n(t) = C_n \exp\left[-\left(\frac{\mu_n \kappa}{L}\right)^2 t\right].$$

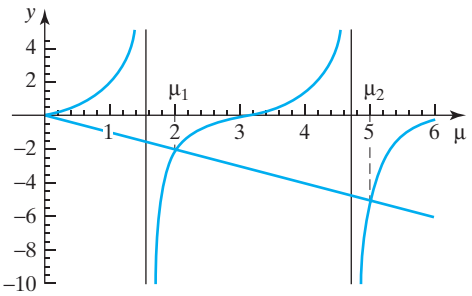


FIGURE 18.24 Graphs of $y = \tan\mu$ and $y = -\mu$ for $\mu > 0$.

The eigenfunction $X_n(x)$ becomes $X_n(x) = B_n \sin\left(\frac{\mu_n x}{L}\right)$, so the eigensolution $X_n T_n$ becomes

$$X_n T_n = D_n \exp\left[-\left(\frac{\mu_n \kappa}{L}\right)^2 t\right] \sin\left(\frac{\mu_n x}{L}\right).$$

We now seek a solution involving the linear combination of eigensolutions

$$u(x,t) = \sum_{n=1}^{\infty} D_n \exp\left[-\left(\frac{\mu_n \kappa}{L}\right)^2 t\right] \sin\left(\frac{\mu_n x}{L}\right),$$

where the constants $D_n = B_n C_n$ are to be determined by use of the initial condition $u(x,0) = \sin(\pi x/L)$. Setting $t = 0$ and using this condition gives

$$\sin\left(\frac{\pi x}{L}\right) = \sum_{n=1}^{\infty} D_n \sin\left(\frac{\mu_n x}{L}\right).$$

Multiplying by $\sin(\mu_m x/L)$ and integrating over $0 \le x \le L$ gives

$$D_n = \left(\frac{2\pi \sin\mu_n}{\pi^2 - \mu_n^2}\right)\left(\frac{p^2 + \mu_n^2}{p(p+1) + \mu_n^2}\right),$$

and so

$$u(x,t) = \sum_{n=1}^{\infty} \left(\frac{2\pi \sin\mu_n}{\pi^2 - \mu_n^2}\right)\left(\frac{p^2 + \mu_n^2}{p(p+1) + \mu_n^2}\right) \exp\left[-\left(\frac{\mu_n \kappa}{L}\right)^2 t\right] \sin\left(\frac{\mu_n x}{L}\right).$$

When obtaining this solution we have used the result

$$\int_0^L \sin(\pi x/L)\sin(\mu_m x/L)dx = \frac{\pi L \sin\mu_m}{\pi^2 - \mu_m^2},$$

and the orthogonality of the eigenfunctions $X_n(x)$ of the associated Sturm–Liouville problem over the interval $0 \le x \le L$ with respect to the weight function $w(x) \equiv 1$, that after integration gives

$$\int_0^L \sin(\mu_m x/L)\sin(\mu_n x/L)dx = \begin{cases} 0, & m \ne n \\ \dfrac{L}{2}\left(\dfrac{p(p+1) + \mu_n^2}{p^2 + \mu_n^2}\right), & m = n, \end{cases}$$

where

$$\sin\mu_n = -\mu_n/\left(p^2+\mu_n^2\right)^{1/2}, \quad \cos\mu_n = p/\left(p^2+\mu_n^2\right)^{1/2}.$$

In the next heat conduction example we consider a problem that requires the use of cylindrical polar coordinates.

EXAMPLE 18.15

a heat problem involving plane polar coordinates

Find the time-dependent temperature distribution $u(r,\theta,t)$ in a thin semicircular metal plate $0 \le r \le 1, 0 \le \theta \le \pi$, given that its plane faces are insulated to prevent heat loss through them, the straight edge of the plate formed by the diameter $0 \le r \le 1, \theta = 0$ and $\theta = \pi$ is insulated, the semicircular boundary is maintained at zero temperature, and the initial temperature distribution is $u(r,\theta,0) = (1-r)\cos\theta$.

Solution The geometry of this problem requires the use of plane polar coordinates, in terms of which the temperature $u(r,\theta,t)$ must satisfy the two-dimensional time-dependent heat equation (see Section 11.6)

$$\frac{\partial u}{\partial t} = \kappa^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2}\right).$$

The bounding diameter $0 \le r \le 1, \theta = 0$, and $\theta = \pi$ is thermally insulated, so as the derivative normal to the diameter is u_θ, the boundary condition on this line becomes $u_\theta(r,0,t) = 0$ and $u_\theta(r,\pi,t) = 0$. The semicircular boundary is maintained at zero temperature, so the boundary condition there is $u(1,\theta,t) = 0$. A routine check shows the initial condition to be appropriate, because it satisfies both the boundary condition on the diameter and the one on the semicircular boundary.

We now separate the variables by seeking elementary solution of the form $u(r,\theta,t) = R(r)\Theta(\theta)T(t)$. Substituting into the heat equation and dividing by $R\Theta T$ gives

$$\frac{T'}{T} = \kappa^2\left(\frac{R''}{R} + \frac{1}{r}\frac{R'}{R} + \frac{1}{r^2}\frac{\Theta''}{\Theta}\right).$$

The expression on the left is only a function of t, and the one on the right is a function of r and θ, so each must be equal to a separation constant. As the temperature must decrease with time, it follows that the separation constant must be negative, so setting it equal to $-\lambda^2$ with $\lambda > 0$, we arrive at the two equations

$$T' + \kappa^2\lambda^2 T = 0 \quad \text{and} \quad r^2\frac{R''}{R} + r\frac{R'}{R} + \lambda^2 r^2 = -\frac{\Theta''}{\Theta}.$$

In the second equation the expression on the left is only a function of r and the one on the right is only a function of θ, so each must be equal to another separation constant μ, so we obtain the two Sturm–Liouville equations

$$\Theta'' + \mu\Theta = 0 \quad \text{and} \quad r^2R'' + rR' + (\lambda^2 r^2 - \mu)R = 0.$$

The general solution for Θ is

$$\Theta(\theta) = A\cos\sqrt{\mu}\,\theta + B\sin\sqrt{\mu}\,\theta,$$

so as the boundary conditions on the diameter are $u_\theta(r,0,t) = 0$ and $u_\theta(r,\pi,t) = 0$, it follows that we must have $\Theta'(\theta)|_{\theta=0} = 0$ and $\Theta'(\theta)|_{\theta=\pi} = 0$. The first of these conditions gives $B = 0$, and the second gives $\sin\sqrt{\mu}\pi = 0$, so $\sqrt{\mu} = 0, 1, \ldots$. Setting $\sqrt{\mu} = m$, and using the fact that the equation for Θ is homogeneous, we may set

the arbitrary constant $A = 1$ when

$$\Theta_m(\theta) = \cos m\theta, \quad \text{for} \quad m = 0, 1, \ldots.$$

how Bessel's equation and its zeros enter into this time-dependent heat equation

The equation for $R(r)$ now becomes Bessel's equation

$$r^2 R'' + r R' + (\lambda^2 r^2 - m^2) R = 0,$$

with the general solution

$$R_m(r) = P J_m(\lambda r) + Q Y_m(\lambda r).$$

The temperature must remain finite throughout the plate, so as $Y_m(\lambda r)$ becomes infinite when $r = 0$, we must set $Q = 0$, reducing the equation to $R_m(r) = J_m(\lambda r)$, where because the equation is homogeneous we have set the arbitrary constant $P = 1$.

To satisfy the boundary condition on the semicircular boundary $u(1, \theta, t) = 0$, we must have $R(1) = 0$, and so λ must satisfy the eigenvalue equation $J_m(\lambda) = 0$, showing that the eigenvalues λ must be the positive zeros $j_{m,n}$ of the Bessel function $J_m(r) = 0$, where $j_{m,n}$ is the nth positive zero of $J_m(r)$. A short list of these zeros can be found in Table 8.1 of Chapter 8.

Using these eigenvalues in the equation for the time variation $T' + \kappa^2 \lambda^2 T = 0$ shows that $T_{m,n}(t) = C_{m,n} \exp\{-j_{m,n}^2 \kappa^2 t\}$, so combining the results for $R(r)$, $\Theta(\theta)$, and $T(t)$, we now seek a solution in the form of the following linear combination of elementary solutions:

$$u(r, \theta, t) = \sum_{m=0,n=1}^{\infty} C_{m,n} J_m(j_{m,n} r) \cos m\theta \exp\{-j_{m,n}^2 \kappa^2 t\}.$$

To find the coefficients $C_{m,n}$ we now make use of the initial condition, the orthogonality of the cosine functions over the interval $0 \le \theta \le \pi$, and the orthogonality of the Bessel functions over the interval $0 \le r \le 1$. Setting $t = 0$ in the preceding series solution and equating the result to the initial condition $u(r, \theta, 0) = (1 - r) \cos\theta$ gives

$$(1 - r) \cos\theta = \sum_{m=0,n=1}^{\infty} C_{m,n} J_m(j_{m,n} r) \cos m\theta.$$

Multiplying this by $\cos\theta$ and integrating over the interval $0 \le \theta \le \pi$ causes every term on the right to vanish, with the exception of the one involving $\cos\theta$ corresponding to $m = 1$. Thus, the required series representation simplifies to

$$(1 - r) \cos\theta = \sum_{n=1}^{\infty} C_{1,n} J_1(j_{1,n} r) \cos\theta,$$

and so after cancellation of the factor $\cos\theta$ to

$$(1 - r) = \sum_{n=1}^{\infty} C_{1,n} J_1(j_{1,n} r).$$

This same result could have been obtained by noticing that as only $\cos\theta$ occurs on the left, the linear independence of cosines of multiple angles requires that all terms involving $\cos m\theta$ on the right must vanish for $m \neq 1$.

To find the coefficients $C_{1,n}$ we multiply the last result by $r J_1(j_{1,s} r)$, integrate over the interval $0 \le r \le 1$, and after using the orthogonality of Bessel functions

derived in (148) of Appendix 2 in Chapter 8, we obtain

$$\int_0^1 r(1-r)J_1(j_{1,s}r)dr = C_{1,s}\frac{1}{2}[J_2(j_{1,s})]^2.$$

Replacing s by n gives

$$C_{1,n} = \frac{2\int_0^1 (r-r^2)J_1(j_{1,n}r)dr}{[J_2(j_{1,n})]^2} \quad \text{for } n = 1, 2, \ldots.$$

In terms of these coefficients $C_{1,n}$, the required solution becomes

$$u(r,\theta,t) = \sum_{n=1}^{\infty} C_{1,n}J_1(j_{1,n}r)\cos\theta \exp\{-j_{1,n}^2\kappa^2 t\}.$$

Evaluating the first few coefficients numerically gives

$$C_{1,1} = 0.917184, \quad C_{1,2} = 0.432800, \quad C_{1,3} = 0.317323, \quad C_{1,4} = 0.232474,$$
$$C_{1,5} = 0.193256, \quad C_{1,6} = 0.158851, \quad C_{1,7} = 0.139139, \quad C_{1,8} = 0.120617.$$

On the diameter bounding the semicircle, when $\theta = 0$ the initial condition is $u(r, 0, t) = 1 - r$, and when $\theta = \pi$ it is $u(r, \pi, t) = r - 1$, so the solution u is discontinuous at $r = 0$ on the bounding diameter.

Figure 18.25 shows a plot of the solution along the insulated diameter as a function of time, using the eight terms in the series solution for $u(r, \theta, t)$ with $\kappa^2 = 0.1$. The plot shows the development of the Gibbs phenomenon at $t = 0$ due to the discontinuity in u at $r = 0$, and the way the temperature along the diameter relaxes to zero as $t \to \infty$. ■

Separation of Variables in the Elliptic Case

Laplace's equation describes many different physical situations, from among which we choose to solve three problems. The first two involve steady-state temperature

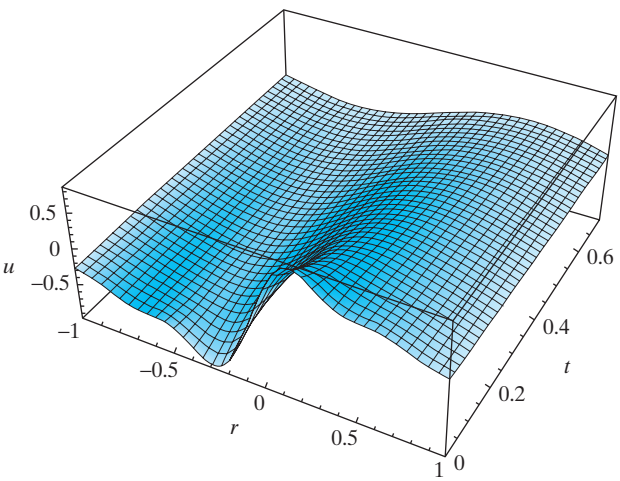


FIGURE 18.25 The relaxation of the initial temperature distribution with time along the diameter bounding the plate.

distributions in two-dimensional regions, and the third involves finding the electrostatic potential distribution inside a spherical cavity. The equation determining the steady-state temperature u in a heat-conducting material is the Laplace equation $\Delta u = 0$, and the first problem to be considered is as follows.

EXAMPLE 18.16

The diagram in Fig. 18.26 shows a rectangular region $0 \le x \le \pi, 0 \le y \le 2$, in which the steady state temperature distribution $u(x, y)$ is required subject to the temperature on the side $0 \le x \le \pi$, $y = 0$, being $u(x, 0) = x \sin x$, and the temperature on the other three sides being maintained at $u = 0$. This can either be considered to represent a cross-section of a long metal bar extending in the z-direction with the boundary conditions on its sides independent of z, or as a thin metal plate with its faces parallel to the (x, y)-plane thermally insulated.

Solution The domain is rectangular with its sides parallel to the coordinate axes, so it is appropriate to express the Laplace equation in terms of the cartesian coordinates x and y so the temperature must satisfy

$$u_{xx} + u_{yy} = 0.$$

Separating variables by setting $u(x, y) = X(x)Y(y)$, substituting into the Laplace equation, dividing by XY, and rearranging terms gives

$$\frac{X''}{X} = -\frac{Y''}{Y}.$$

As the expression on the left is a function of only x and the one on the right is a function of only y, these expressions must be equal to a separation constant k, so that

$$\frac{X''}{X} = -\frac{Y''}{Y} = k.$$

The sign of the separation constant must be chosen so the boundary conditions are satisfied. As $u(x, y) = X(x)Y(y)$, and neither $X(x)$ nor $Y(y)$ can be identically zero, the boundary conditions $u(0, y) = 0$ and $u(\pi, y) = 0$ imply that $X(0) = X(\pi) = 0$. When $k > 0$, the general solution for $X(x)$ is $X(x) = A \cosh x\sqrt{k} + B \sinh x\sqrt{k}$, and the boundary conditions can only be satisfied if $A = B = 0$, which is impossible. Consequently, k must be negative, so we set $k = -\lambda^2$, where λ is positive and real. The separated equations give the following Sturm–Liouville equation

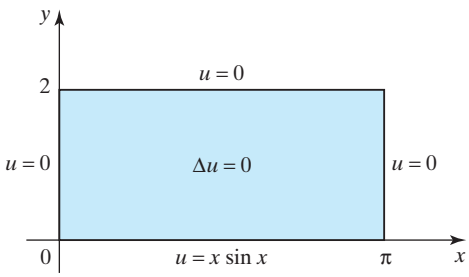


FIGURE 18.26 The rectangular region and its boundary conditions.

for $X(x)$ and the equation for $Y(y)$:

$$X'' + \lambda^2 X = 0 \quad \text{and} \quad Y'' - \lambda^2 Y = 0.$$

Solving for X gives

$$X(x) = \widetilde{A}\cos\lambda x + \widetilde{B}\sin\lambda x,$$

and imposing the left boundary condition $X(0) = 0$ shows that $\widetilde{A} = 0$. The imposition of the right boundary condition $X(\pi) = 0$ gives $\widetilde{B}\sin\lambda\pi = 0$, so as $\widetilde{B} \neq 0$, it follows that the eigenvalues are the zeros of $\sin\pi x$, and so

$$\lambda_n = n, \quad \text{for } n = 1, 2, \ldots.$$

Thus, the eigenfunctions are proportional to

$$X_n(x) = \widetilde{B}\sin nx, \quad \text{for } n = 1, 2, \ldots,$$

where, as the equation for $X_n(x)$ is homogeneous, the value of $\widetilde{B}$ is unimportant.

Solving the differential equation for $Y(y)$ gives

$$Y_n(y) = \widetilde{C}\cosh ny + \widetilde{D}\sinh ny.$$

The boundary condition $u(x, 2) = 0$ is equivalent to $X(x)Y(2) = 0$, but as $X(x)$ is not identically zero, we must have $Y(2) = 0$. Applying this condition to $Y_n(y)$ gives

$$0 = \widetilde{C}\cosh 2n + \widetilde{D}\sinh 2n,$$

but only the ratio is important, so we can set $\widetilde{D} = 1$ when

$$\widetilde{C} = -\frac{\sinh 2n}{\cosh 2n}.$$

Using this result in the expression for $Y_n(y)$ gives

$$Y_n(y) = \frac{\widetilde{C}}{\cosh 2n}(\sinh ny\cosh 2n - \cosh ny\sinh 2n) = \frac{\widetilde{C}}{\cosh 2n}\sinh n(y-2).$$

If we replace the product $\widetilde{B}\widetilde{C}/\cosh 2n$ by C_n, the eigensolution $u_n(x, y) = X_n(x)Y_n(y)$ becomes

$$u_n(x, y) = C_n \sin nx \sinh n(y-2), \quad \text{for } n = 1, 2, \ldots.$$

We now seek a solution of the boundary value problem in the form of the linear combination of the eigensolutions

$$u(x, y) = \sum_{n=1}^{\infty} u_n(x, y) = \sum_{n=1}^{\infty} C_n \sin nx \sinh n(y-2).$$

To determine the coefficients C_n we must use the boundary condition $u(x, 0) = x\sin x$ together with the orthogonality properties of the set of functions $\{\sin nx\}|_1^\infty$ over the interval $0 \le x \le \pi$. Setting $y = 0$ in $u(x, y)$ and multiplying the result by $\sin mx$, integrating over $0 \le x \le \pi$, and using the orthogonality properties of the set of sine functions gives

$$\int_0^\pi x\sin x\sin nx\,dx = -C_n\sinh 2n\int_0^\pi \sin^2 nx\,dx, \quad n = 1, 2, \ldots.$$

Evaluating the integrals and solving for C_n we find that

$$C_1 = -\frac{\pi}{2\sinh 2} \quad \text{and} \quad C_n = \frac{4n(1+(-1)^n)}{(n^2-1)^2\pi\sinh 2n} \quad \text{for } n = 2, 3, \ldots.$$

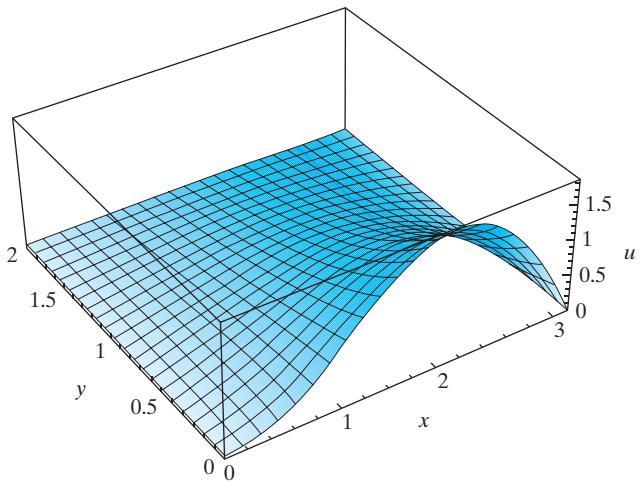


FIGURE 18.27 A plot of the temperature distribution $u(x, y)$ using five terms.

The problem is solved by substituting these values of C_n into

$$u(x, y) = \sum_{n=1}^{\infty} C_n \sin nx \sinh n(y - 2).$$

Figure 18.27 shows a computer plot of the temperature distribution $u(x, y)$ in the region $0 \le x \le \pi$, $0 \le y \le 2$ obtained by using the preceding result with five terms. ■

The following is another example of the application of the method of separation of variables to the Laplace equation when finding the steady state temperature distribution.

EXAMPLE 18.17

Find the steady state temperature distribution in the semicircular region of radius ρ lying in the upper half-plane and centered on the origin, as shown in Fig. 18.28. The temperature on the straight boundary is $u = 0$, and that on the semicircular boundary is $u = u_0\theta(\pi - \theta)$.

Solution The geometry of the problem suggests that the Laplace equation for the steady state temperature distribution u should be expressed in terms of the polar coordinates r and θ. In terms of these variables the Laplace equation $\Delta u = 0$ becomes

$$u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0.$$

To separate the variables we now set $u(r, \theta) = R(r)\Theta(\theta)$ and substitute into the equation. After dividing by $R\Theta$ and rearranging terms, we find that

$$r^2\frac{R''}{R} + r\frac{R'}{r} = -\frac{\Theta''}{\Theta},$$

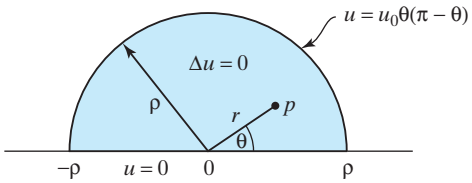


FIGURE 18.28 The semicircular domain and its boundary conditions.

but as the expression on the left is a function of only r and the one on the right is a function of only θ, both must be equal to a separation constant k, so we have

$$r^2R'' + rR' - kR = 0 \quad \text{and} \quad \Theta'' + k\Theta = 0.$$

The sign of k is determined by the fact that only when $k > 0$ will the θ variation be periodic in nature, as would be expected, because increasing θ by a multiple of 2π will simply reproduce the original problem. If we set $k = \lambda^2$, the functions R and Θ are seen to satisfy the two equations

$$r^2R'' + rR' - \lambda^2R = 0 \quad \text{and} \quad \Theta'' + \lambda^2\Theta = 0.$$

how the Cauchy–Euler equation arises

The first of these equations is a **Cauchy–Euler equation**, which was seen in Section 6.5 to have the general solution

$$R(r) = \widetilde{A}r^{\lambda} + \widetilde{B}\frac{1}{r^{\lambda}}.$$

As the solution must be *finite* at the origin, we must set $\widetilde{B} = 0$, so $R(r)$ must be of the form $R(r) = \widetilde{A}r^{\lambda}$. Now, as $u(r, \theta) = R(r)\Theta(\theta)$ and $u(r, 0) = u(r, \pi) = 0$ (in polar coordinates these two conditions represent the boundary condition on the straight line boundary), it follows that the boundary conditions for Θ are $\Theta(0) = \Theta(\pi) = 0$.

The general solution for Θ is

$$\Theta(\theta) = \widetilde{C}\cos\lambda\theta + \widetilde{D}\sin\lambda\theta,$$

so imposing the first of the boundary conditions gives $\widetilde{C} = 0$, and when the second one is imposed we find that λ must satisfy

$$0 = \widetilde{D}\sin\lambda\pi,$$

so the eigenvalues λ_n are

$$\lambda_n = n, \quad \text{for } n = 1, 2, \ldots.$$

The eigenfunctions $R_n(r)$ become

$$R_n(x) = A_nr^n, \quad \text{for } n = 1, 2, \ldots,$$

and the eigensolutions $u_n(r, \theta) = A_nr^n \sin n\theta$, where the product of the arbitrary constants $\widetilde{A}\widetilde{D}$, each of which depends on n, has been denoted by A_n.

We now seek a solution in the form of the linear combination of the eigensolutions

$$u(r, \theta) = \sum_{n=1}^{\infty} u_n(r, \theta) = \sum_{n=1}^{\infty} A_nr^n \sin n\theta.$$

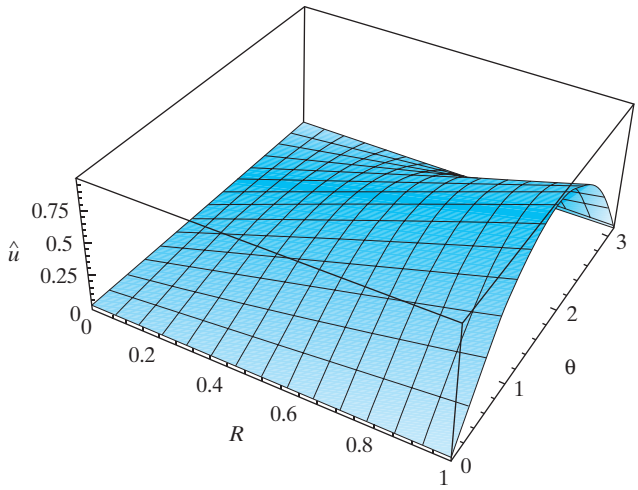


FIGURE 18.29 A plot of the normalized solution $\hat{u} = (\pi/8u_0)u(r, \theta)$.

Substituting the boundary condition $u(\rho, \theta) = u_0\theta(\pi - \theta)$ on the left of this series and setting $r = \rho$ in the expression on the right gives

$$u_0\theta(\pi - \theta) = \sum_{n=1}^{\infty} A_n \rho^n \sin n\theta.$$

The coefficients A_n now follow from the orthogonality properties of the sine function over the interval $0 \le \theta \le \pi$. Multiplying the last result by $\sin m\theta$ and integrating over the interval $0 \le \theta \le \pi$, we find that

$$2u_0\left(\frac{1 - (-1)^n}{n^3}\right) = \frac{1}{2}A_n\rho^n\pi \quad \text{and so} \quad A_n = \frac{4u_0}{\pi}\frac{(1 - (-1)^n)}{n^3\rho^n}.$$

Substituting these coefficients into the series now gives the required solution,

$$u(r, \theta) = \frac{8u_0}{\pi}\sum_{n=1}^{\infty}\left(\frac{r}{\rho}\right)^{2n-1}\frac{\sin(2n-1)\theta}{(2n-1)^3}.$$

Figure 18.29 shows a plot of $\hat{u} = (\pi/8u_0)u(r, \theta)$ as a function of $R = r/\rho$ for $0 \le R \le 1$ and $0 \le \theta \le \pi$ using 10 terms of the series. ■

The next example involving Laplace's equation is a three-dimensional problem for which spherical polar coordinates form the natural coordinate system to be used. This example also shows how Legendre polynomials arise naturally when we work with Laplace's equation in spherical polar coordinates.

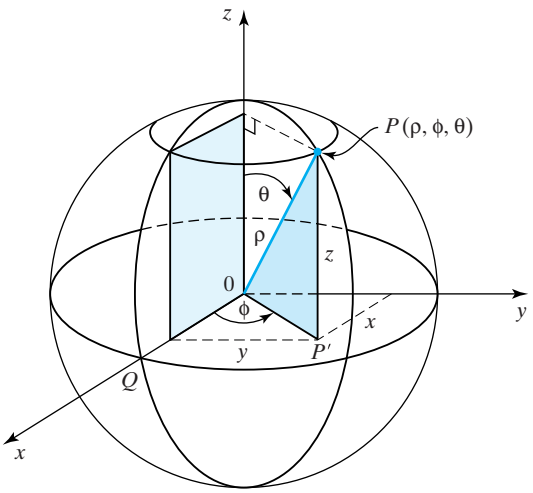


FIGURE 18.30 The spherical polar coordinate system.

EXAMPLE 18.18

a problem involving spherical polar coordinates

Find the electrostatic potential inside a spherical cavity of radius ρ when the bottom half of the spherical boundary is maintained at a potential U_0 and the upper half is maintained at a potential U_1.

Solution The geometry of the problem indicates that for simplicity spherical polar coordinates should be used, because the boundary of the region involved is a sphere of radius ρ. Figure 18.30 shows the standard system of spherical coordinates. As the potential on the boundary assumes a different constant value on each of two hemispheres, the problem will be simplified if the origin is taken to be at the center of the sphere with the z-axis chosen so the potential is $u = U_1$ on the upper hemisphere where $z > 0$, corresponding to $r = \rho$ and $0 \le \theta < \frac{\pi}{2}$, and $u = U_0$ on the lower hemisphere where $z < 0$, corresponding to $r = \rho$ and $\frac{\pi}{2} < \theta \le \pi$.

In this case the boundary conditions are such that there is no variation with respect to the angle ϕ (called the *azimuthal* angle), so as the potential inside the spherical cavity will depend only on r and θ we set $u = u(r, \theta)$. Making use of the expression for the Laplacian in spherical polar coordinates found in Example 11.23(b) of Chapter 11, and setting the partial derivative with respect to ϕ equal to zero, because there is no variation with respect to ϕ, gives

$$\Delta u = \frac{1}{r^2 \sin\theta}\left[\frac{\partial}{\partial r}\left(r^2 \sin\theta \frac{\partial u}{\partial r}\right) + \frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u}{\partial \theta}\right)\right] = 0,$$

or

$$r^2 \frac{\partial^2 u}{\partial r^2} + 2r\frac{\partial u}{\partial r} + \cot\theta \frac{\partial u}{\partial \theta} + \frac{\partial^2 u}{\partial \theta^2} = 0.$$

For what is to follow, derivatives with respect to θ need to be transformed into derivatives with respect to ξ, where $\xi = \cos\theta$. Using the results obtained from the chain rule

$$\frac{\partial u}{\partial \theta} = -\sin\theta \frac{\partial u}{\partial \xi} \quad \text{and} \quad \frac{\partial^2 u}{\partial \theta^2} = \sin^2\theta \frac{\partial^2 u}{\partial \xi^2} - \cos\theta \frac{\partial u}{\partial \xi},$$

we find that

$$\Delta u = r^2 \frac{\partial^2 u}{\partial r^2} + 2r \frac{\partial u}{\partial r} - 2\xi \frac{\partial u}{\partial \xi} + (1 - \xi^2) \frac{\partial^2 u}{\partial \xi^2} = 0.$$

Separating variables by seeking elementary solutions of the form $u(r, \xi) = R(r)Q(\xi)$, substituting into the preceding equation, and then dividing by RQ gives

$$\frac{r^2 R'' + 2r R'}{R} = \frac{2\xi Q' - (1 - \xi^2)Q''}{Q} = k,$$

where, as $R = R(r)$ and $Q = Q(\xi)$, these expressions must both be equal to a separation constant k whose value will be assigned later. Now that the variables have been separated, the two differential equations that follow from this are

$$r^2 R'' + 2rR' - kR = 0 \quad \text{and} \quad (1 - \xi^2)Q'' - 2\xi Q' + kQ = 0.$$

If we now choose the separation constant to be $k = n(n + 1)$ with $n = 0, 1, \ldots,$ the second equation becomes

$$(1 - \xi^2)Q'' - 2\xi Q' + n(n + 1)Q = 0,$$

and from Section 8.2 of Chapter 8 its solution is seen to be $Q(\xi) = P_n(\xi)$, where $P_n(\xi)$ is the Legendre polynomial of degree n. The equation for R now becomes the Cauchy–Euler equation

$$r^2 R'' + 2rR' - n(n + 1)R = 0.$$

The solution of this equation is found by setting $R = r^\alpha$ and solving for α. As a result we find $\alpha = n$ or $\alpha = -(n + 1)$, so the general solution for $R(r)$ is

$$R(r) = Ar^n + Br^{-(n+1)}.$$

The potential $u(r, \xi)$ must remain finite at the origin, so we must set $B = 0$. Thus, the required elementary eigensolution $u_n(r, \xi) = R(r)Q(\xi)$ becomes

$$u_n(r, \xi) = A_n r^n P_n(\xi).$$

how a Fourier–Legendre expansion arises

We now use this result to find the potential inside the sphere in the form of the linear combination of eigensolutions

$$u(r, \xi) = \sum_{n=0}^{\infty} A_n r^n P_n(\xi),$$

which form a Fourier–Legendre expansion of $u(r, \xi)$.

In terms of the new variable ξ, the boundary conditions on the spherical boundary $r = \rho$ become $u(\rho, \xi) = U_0$ for $-1 \le \xi < 0$ and $u(\rho, \xi) = U_1$ for $0 < \xi \le 1$. The coefficients A_n now follow by setting $r = \rho$ in the Fourier–Legendre expansion for $u(r, \xi)$, substituting the boundary conditions, multiplying by $P_m(\xi)$, and integrating the result with respect to ξ over the interval $-1 \le \xi \le 1$, followed by use of the orthogonality property of Legendre polynomials (see Chapter 8),

$$\int_{-1}^{1} P_m(\xi) P_n(\xi) d\xi = \begin{cases} \frac{2}{2n+1}, & m = n \\ 0, & m \ne n. \end{cases}$$

When this is done the coefficients A_n are found to be given by

$$A_n = \left(\frac{2n + 1}{2\rho^n}\right)\left(\int_{-1}^{0} U_0 P_n(\xi) d\xi + \int_{0}^{1} U_1 P_n(\xi) d\xi\right), \quad \text{for } n = 0, 1, \ldots,$$

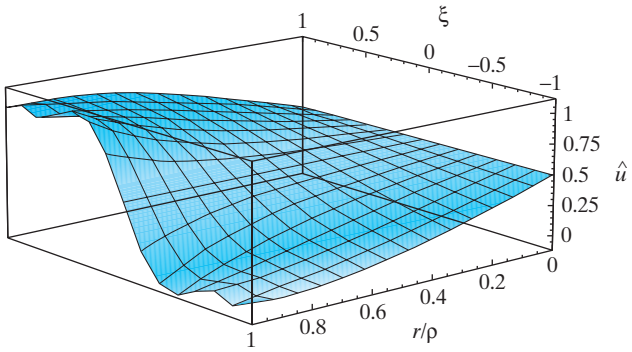


FIGURE 18.31 A plot of the normalized solution $\hat{u}(r, \xi)$.

and so

$$A_0 = \frac{1}{2}(U_0 + U_1),\ A_1 = \frac{3}{4\rho}(U_1 - U_0),\ A_2 = 0,\ A_3 = -\frac{7}{16\rho^3}(U_1 - U_0),\ A_4 = 0$$

$$A_5 = \frac{11}{32\rho^5}(U_1 - U_0),\ A_6 = 0,\ A_7 = -\frac{55}{256\rho^7}(U_1 - U_0),\ A_8 = 0, \ldots.$$

Substituting for the A_n in the Fourier–Legendre series for $u(r, \xi)$ shows the solution to be

$$\frac{u(r, \xi) - U_0}{U_1 - U_0} =$$

$$\left\{\frac{1}{2} + \frac{3}{4}\left(\frac{r}{\rho}\right)P_1(\xi) - \frac{7}{16}\left(\frac{r}{\rho}\right)^3 P_3(\xi) + \frac{11}{32}\left(\frac{r}{\rho}\right)^5 P_5(\xi) - \frac{55}{256}\left(\frac{r}{\rho}\right)^7 P_7(\xi) + \cdots\right\},$$

for $-1 \le \xi \le 1$ with $\xi = \cos\theta$.

Figure 18.31 shows a plot of $\hat{u}(r, \xi) = [u(r, \xi) - U_0]/(U_1 - U_0)$ obtained using the preceding approximation with $0 \le r/\rho \le 1$ and $-1 \le \xi \le 1$. The plot exhibits the start of the Gibbs phenomenon in this Fourier–Legendre expansion due to the discontinuity in the boundary condition across $r = \rho$ when $\theta = \pi/2$. ■

So far the method of separation of variables has only been applied to homogeneous equations. The next example illustrates a way in which the nonhomogeneous one-dimensional heat equation may be solved by using variation of parameters in the method of separation of variables.

EXAMPLE 18.19

The temperature $u(x, t)$ in a slab of metal $0 < x < L$ with heat generated in it at time t and position x at a rate $H(x, t)$ is determined by the nonhomogeneous heat equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2} + H(x, t),$$

subject to the initial condition

$$u(x, 0) = U(x)$$

and the boundary conditions

$$u(0, t) = u(L, t) = 0 \text{ for } t > 0.$$

Find the temperature distribution $u(x, t)$ in the slab by combining method of variation of parameters with separation of the variables.

Solution The nonhomogeneous term does not allow separation of variables to be used directly, so a modified approach must be adopted. Let us consider first the solution of the problem when $H(x, t) \equiv 0$.

Separating variables by setting $u(x, t) = X(x)T(t)$ and proceeding in the usual manner leads to the separated equations

$$\frac{T'(t)}{\kappa T(t)} = \frac{X''(x)}{X(x)}.$$

Introducing a separation constant $-\lambda$ with $\lambda > 0$, where the negative sign is chosen to make the solution satisfy the physical requirement that it decays with time, we arrive at the two separated ordinary differential equations

$$\frac{dT}{dt} = -\lambda\kappa T \quad \text{and} \quad \frac{d^2X}{dx^2} + \lambda X = 0.$$

To satisfy the boundary conditions on the temperature $u(x, t)$, the function $X(x)$ must satisfy the boundary conditions $X(0) = X(L) = 0$. The equation for $X(x)$ together with these boundary conditions is a Sturm–Liouville problem that determines the eigenvalues λ_n and the associated eigenfunctions $X_n(x)$. As the general solution for $X(x)$ is

$$X(x) = A\cos(\sqrt{\lambda}x) + B\sin(\sqrt{\lambda}x),$$

the boundary conditions will only be satisfied when $\lambda = (n\pi/L)^2$ and $A = 0$, so the eigenvalues are $\lambda_n = (n\pi/L)^2$ and the associated eigenfunctions can be taken to be $X_n(x) = \sin(n\pi x/L)$, with $n = 1, 2 \ldots$.

Integrating the equation for the time variation $T(t)$ with $\lambda = \lambda_n$ gives $T_n(t) = \exp(-\lambda_n\kappa t)$, so the elementary solutions for this problem are

$$u_n(x, t) = \exp\left(-\left(\frac{n\pi}{L}\right)^2 \kappa t\right) \sin\left(\frac{n\pi x}{L}\right), \text{ with } n = 1, 2, \ldots.$$

It follows from this that the solution for the temperature distribution will be of the form

$$u(x, t) = \sum_{n=1}^{\infty} a_n u_n(x, t) = \sum_{n=1}^{\infty} a_n \exp\left(-\left(\frac{n\pi}{L}\right)^2 \kappa t\right) \sin\left(\frac{n\pi x}{L}\right).$$

The coefficients a_n follow in the usual manner by setting $t = 0$ and using the initial condition that $n(x, 0) = U(x)$, when we find that

$$U(x) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right).$$

Multiplying this result by $\sin(n\pi x/L)$ and integrating over the interval $0 \le x \le L$ gives

$$a_n = \frac{2}{L}\int_0^L U(x)\sin\left(\frac{n\pi x}{L}\right)dx, \quad \text{for } n = 1, 2, \ldots.$$

This completes the solution for the temperature $u(x,t)$ when the heat equation is homogeneous, because

$$u(x,t) = \sum_{n=1}^{\infty} a_n u_n(x,t) = \sum_{n=1}^{\infty} a_n \exp\left(-\left(\frac{n\pi}{L}\right)^2 \kappa t\right) \sin\left(\frac{n\pi x}{L}\right).$$

To make use of this solution in the nonhomogeneous case, we start by seeking a solution of the form

$$u(x,t) = \sum_{n=1}^{\infty} \Psi_n(t) \sin\left(\frac{n\pi x}{L}\right),$$

where the functions $\Psi_n(t)$ are still to be determined. We then expand $H(x,t)$ in terms of x as

$$H(x,t) = \sum_{n=1}^{\infty} H_n(t) \sin\left(\frac{n\pi x}{L}\right),$$

where the time-dependent coefficients $H_n(t)$ are obtained from $H(x,t)$ by multiplying this last result by $\sin(n\pi x/L)$ and integrating over the interval $0 \le x \le L$.

The initial condition $u(x,0) = U(x)$ has already been expanded as

$$U(x) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right), \text{ with } a_n = \frac{2}{L}\int_0^L U(x) \sin\left(\frac{n\pi x}{L}\right) dx, \quad \text{for } n = 1, 2, \ldots$$

so after substituting these results in the PDE and combining terms in $\sin(n\pi x/L)$, we obtain

$$\sum_{n=1}^{\infty} \left[\frac{d\Psi_n(t)}{dt} + \kappa\left(\frac{n\pi}{L}\right)^2 \Psi_n(t) - H_n(t)\right] \sin\left(\frac{n\pi x}{L}\right) = 0.$$

As the right-hand side of this equation is zero, multiplying the series by $\sin(n\pi x/L)$ and integrating the result over the interval $0 \le x \le L$ shows that the unknown functions $\Psi_n(t)$ are solutions of the linear first order equation

$$\frac{d\Psi_n(t)}{dt} + \kappa\left(\frac{n\pi}{L}\right)^2 \Psi_n(t) = H_n(t), \text{ with } n = 1, 2, \ldots.$$

The initial conditions for these equations follow from the two different expressions for $u(x,0)$, namely,

$$u(x,0) = \sum_{n=1}^{\infty} \Psi_n(0) \sin\left(\frac{n\pi x}{L}\right) \quad \text{and} \quad u(x,0) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right).$$

These must be true for all x, so when equated they give $\Psi_n(0) = a_n$, for $n = 1, 2, \ldots$. A straightforward integration of the linear first order differential equations for $\Psi_n(t)$

shows the solutions, subject to these initial conditions, to be

$$\Psi_n(t) = a_n \exp\left(-\left(\frac{n\pi}{L}\right)^2 \kappa t\right) + \int_0^t \exp\left(-\left(\frac{n\pi}{L}\right)^2 \kappa(t-s)\right) H_n(s)ds, \text{ for } n = 1, 2, \ldots.$$

Finally, after substituting for $\Psi_n(t)$ in

$$u(x,t) = \sum_{n=1}^{\infty} \Psi_n(t) \sin\left(\frac{n\pi x}{L}\right),$$

we arrive at the required solution

$$u(x,t) = \sum_{n=1}^{\infty} a_n \exp\left(-\left(\frac{n\pi}{L}\right)^2 \kappa t\right) \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} \left(\int_0^t \exp\left(-\left(\frac{n\pi}{L}\right)^2 \kappa(t-s)\right) H_n(s)ds\right) \sin\left(\frac{n\pi x}{L}\right).$$

The first summation on the right is seen to be the solution of the homogeneous equation, whereas the second summation represents the contribution made to the solution by the nonhomogeneous term. ■

The following example shows how the wave equation can be solved by separation of variables when the boundary conditions are dependent on the time.

EXAMPLE 18.20

Solve the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

in the interval $0 \le x \le L$, subject to the initial conditions

$$u(x,0) = f(x) \quad \text{and} \quad u_t(x,0) = g(x)$$

and the time-dependent boundary conditions

$$u(0,t) = h(t) \quad \text{and} \quad u(L,t) = k(t).$$

Solution To take account of the time-dependent boundary conditions, we define an auxiliary function

$$v(x,t) = \left(\frac{L-x}{L}\right) h(t) + \left(\frac{x}{L}\right) k(t)$$

that agrees with the boundary conditions at $x = 0$ and $x = L$. Next we seek a solution $u(x,t)$ of the form

$$u(x,t) = v(x,t) + w(x,t).$$

With this choice of $u(x,t)$, it is seen that $w(x,t)$ must be a solution of

$$\frac{\partial^2 w}{\partial t^2} = c^2 \frac{\partial^2 w}{\partial x^2} + \left(\frac{x-L}{L}\right) \frac{d^2 h}{dt^2} - \left(\frac{x}{L}\right) \frac{d^2 k}{dt^2},$$

with

$$w(x,0) = f(x) + \left(\frac{x-L}{L}\right)h(0) - \left(\frac{x}{L}\right)k(0) = F(x), \text{ say,}$$
$$w_t(x,0) = g(x) + \left(\frac{x-L}{L}\right)h'(0) - \left(\frac{x}{L}\right)k'(0) = G(x), \text{ say,}$$

and

$$w(0,t) = w(L,t) = 0.$$

The trick now is to write $w(x,t) = P(x,t) + Q(x,t)$, with $P(x,t)$ the solution of the homogeneous boundary value problem

$$\frac{\partial^2 P}{\partial t^2} = c^2 \frac{\partial^2 P}{\partial x^2},$$

with the initial conditions

$$P(x,0) = F(x),\ P_t(x,0) = G(x)$$

and the homogeneous boundary conditions

$$P(0,t) = P(L,t) = 0.$$

Arguments similar to those used with Example 18.11 then show that

$$P(x,t) = \sum_{n=1}^{\infty}\left[A_n \cos\left(\frac{n\pi ct}{L}\right) + B_n \sin\left(\frac{n\pi ct}{L}\right)\right]\sin\left(\frac{n\pi x}{L}\right),$$

where

$$A_n = \frac{2}{L}\int_0^L F(x)\sin\left(\frac{n\pi x}{L}\right)dx \text{ and } B_n = \frac{2}{n\pi c}\int_0^L G(x)\sin\left(\frac{n\pi x}{L}\right)dx, n = 1, 2, \ldots.$$

The function $Q(x,t)$ is then a solution of the nonhomogeneous problem

$$\frac{\partial^2 Q}{\partial t^2} = c^2\frac{\partial^2 Q}{\partial x^2} + \left(\frac{x-L}{L}\right)\frac{d^2h}{dt^2} - \left(\frac{x}{L}\right)\frac{d^2k}{dt^2}.$$

If we use the method of Example 18.19, the solution $Q(x,t)$ becomes

$$Q(x,t) = \sum_{n=1}^{\infty}\Psi_n(t)\sin\left(\frac{n\pi x}{L}\right),$$

where

$$\Psi_n(t) = \frac{L}{n\pi}\int_0^t \sin\left(\frac{n\pi}{L}(t-\tau)\right)S_n(\tau)d\tau,$$

with

$$S_n(t) = \frac{2}{L}\int_0^L\left[\left(\frac{x-L}{L}\right)\frac{d^2h}{dt^2} - \left(\frac{x}{L}\right)\frac{d^2k}{dt^2}\right]\sin\left(\frac{n\pi x}{L}\right)dx.$$

The next example concerns the Laplace equation subject to Dirichlet conditions that are imposed on the boundaries of an annulus, and it demonstrates how a logarithmic term can appear in the solution.

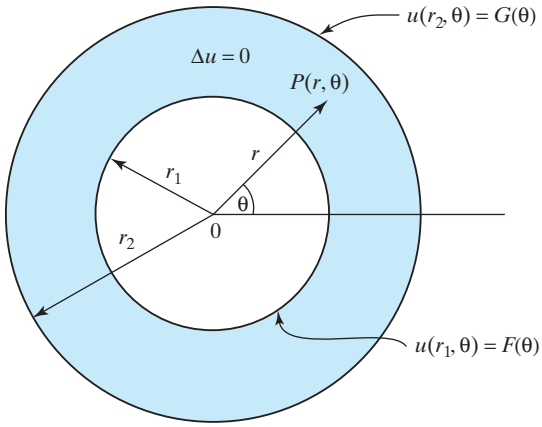


FIGURE 18.32 The Laplace equation in the annulus $r_1 \le r \le r_2$.

EXAMPLE 18.21 Find solution $u(r, \theta)$ of the Laplace equation in cylindrical polar coordinates

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0,$$

in the annulus $r_1 \le r \le r_2$ shown in Fig. 18.32, where $u(r, \theta)$ is periodic in θ with period 2π and subject to the general Dirichlet boundary conditions

$$u(r_1, \theta) = F(\theta) \quad \text{and} \quad u(r_2, \theta) = G(\theta),$$

where $F(\theta)$ and $G(\theta)$ are continuous functions of θ that are periodic with period 2π. Apply the result to find $u(r, \theta)$ in the annulus $2 \le r \le 3$, when $F(\theta) = 1 + \sin\theta$ and $G(\theta) = \cos\theta + \frac{1}{3}\cos 2\theta$.

Solution First, it is necessary to remember that in polar coordinates the polar angle θ is indeterminate to within a multiple of 2π, so for $u(r, \theta)$ to be a continuous function of θ it is necessary that the Dirichlet (boundary) conditions should be periodic with period 2π. This can be expressed analytically by requiring that $F(\theta) = F(\theta + 2\pi)$ and $G(\theta) = G(\theta + 2\pi)$.

Separating variables by writing $u(r, \theta) = R(r)\Theta(\theta)$, substituting $u(r, \theta)$ into the Laplace equation, dividing by $R(r)\Theta(\theta)$, and separating the terms in $R(r)$ and $\Theta(\theta)$ gives

$$\frac{r^2 R''(r) + rR'(r)}{R(r)} = -\frac{\Theta''(\theta)}{\Theta(\theta)} = \lambda,$$

where λ is a separation constant whose values and sign remain to be determined.

The equation for $\Theta(\theta)$, namely $\Theta'' + \lambda\Theta = 0$, will only be periodic in θ when $\lambda > 0$, and it will only be periodic with period 2π if $\lambda = n^2$ with $n = 0, 1, \ldots$. Thus, the eigenvalues of the problem are $\lambda_n = n^2$ and the associated eigenfunctions are

$$\Theta_n(\theta) = A_n \cos(n\theta) + B_n \sin(n\theta), \quad \text{for } n = 0, 1, \ldots.$$

Setting $\lambda = \lambda_n$ in the equation for $R(r)$ shows that it must be a solution of the Cauchy–Euler equation

$$r^2 \frac{d^2 R}{dr^2} + r \frac{dR}{dr} + n^2 R = 0.$$

When $n = 0$, cancelling r, setting $dR/dr = v(r)$, separating variables, and solving for v gives $v = b_0/r$, with b_0 an arbitrary constant of integration. After we replace $v(r)$ by dR/dr in this last result, a further integration gives

$$R_0(r) = a_0 + b_0 \ln r,$$

with a_0 as a second arbitrary constant of integration. When $n = 1, 2, \ldots,$ the Cauchy–Euler equation has the usual solution

$$R_n(r) = a_n r^n + \frac{b_n}{r^n},$$

with a_n and b_n arbitrary constants.

Adding these results, which is permissible because Laplace's equation is linear and homogeneous, shows that we must now seek a solution for $u(r, \theta)$ of the form

$$u(r, \theta) = a_0 + b_0 \ln r + \sum_{n=1}^{\infty} \left[\left(a_n r^n + \frac{b_n}{r^n} \right) \cos(n\theta) + \left(c_n r^n + \frac{d_n}{r^n} \right) \sin(n\theta) \right],$$

though at present it is unclear how the coefficients $a_n, b_n, c_n,$ and d_n are to be determined.

The approach we now use to find these coefficients in the series for $u(r, \theta)$ involves first expanding the Dirichlet condition $F(\theta)$ as Fourier series in θ over the interval $0 \le \theta \le 2\pi$ (remember that $F(\theta)$ is periodic in θ with period 2π). Then, after setting $r = r_1$ in the expression for $u(r, \theta)$ and using the Dirichlet boundary condition $u(r_1, \theta) = F(\theta)$, we will equate the known coefficients of $\cos(n\theta)$ and $\sin(n\theta)$ in the expansion of $F(\theta)$ and the unknown coefficients of the corresponding terms in $\cos(n\theta)$ and $\sin(n\theta)$ in the representation of $u(r_1, \theta)$. A further set of equations will then be obtained in similar fashion by expanding $G(\theta)$ as a Fourier series, setting $r = r_2$ in $u(r, \theta)$, and using the second Dirichlet boundary condition, which gives $u(r_2, \theta) = G(\theta)$. Taken together, these equations will determine all of the coefficients $a_n, b_n, c_n,$ and d_n.

Accordingly, let us represent the Fourier series expansions of $F(\theta)$ and $G(\theta)$ as follows:

$$F(\theta) = \tfrac{1}{2} P_0 + \sum_{n=1}^{\infty} [P_n \cos(n\theta) + Q_n \sin(n\theta)]$$

and

$$G(\theta) = \tfrac{1}{2} S_0 + \sum_{n=1}^{\infty} [S_n \cos(n\theta) + T_n \sin(n\theta)].$$

Equating the coefficients of corresponding terms in $\cos(n\theta)$ and $\sin(n\theta)$ gives

$$\begin{aligned} \tfrac{1}{2} P_0 &= a_0 + b_0 \ln r_1 & \tfrac{1}{2} S_0 &= a_0 + b_0 \ln r_2 \\ P_n &= a_n r_1^n + \frac{b_n}{r_1^n} & Q_n &= c_n r_1^n + \frac{d_n}{r_1^n} \\ S_n &= a_n r_2^n + \frac{b_n}{r_2^n} & T_n &= c_n r_2^n + \frac{d_n}{r_2^n}. \end{aligned}$$

Once these equations have been solved for a_n, b_n, c_n, and d_n, the expansion of $u(r, \theta)$ can be determined, so the general approach to the solution of the Dirichlet problem for Laplace's equation in an annulus has been established.

When $F(\theta) = 1 + \sin\theta$ and $G(\theta) = \cos\theta + \frac{1}{3}\cos 2\theta$, the solution simplifies, because the functions $F(\theta)$ and $G(\theta)$ are already their own Fourier series. The only nonzero coefficients in the Fourier expansion of $F(\theta)$ and $P_0 = 2$ and $Q_1 = 1$, whereas the only nonzero coefficients in the Fourier expansion of $G(\theta)$ are $S_1 = 1$ and $S_2 = \frac{1}{3}$. Consequently, we only need equate coefficients of terms up to the multiple 2θ, so that when $r = 2$ we obtain

$$1 = a_0 + b_0 \ln 2, \quad 0 = 2a_1 + \tfrac{1}{2}b_1, \quad 0 = 4a_2 + \tfrac{1}{4}b_2,$$
$$1 = 2c_1 + \tfrac{1}{2}d_1, \quad 0 = 4c_2 + \tfrac{1}{4}d_2,$$

and when $r = 3$ we obtain

$$0 = a_0 + b_0 \ln 3, \quad 1 = 3a_1 + \tfrac{1}{3}b_1, \quad \tfrac{1}{3} = 9a_2 + \tfrac{1}{9}b_2,$$
$$0 = 3c_1 + \tfrac{1}{3}d_1, \quad 0 = 9c_2 + \tfrac{1}{9}d_2.$$

These have the solutions

$$a_0 = -\frac{\ln 3}{\ln(2/3)}, \; b_0 = \frac{1}{\ln(2/3)}, \; a_1 = \frac{3}{5}, \; b_1 = -\frac{12}{5}, \; a_2 = \frac{3}{65},$$
$$b_2 = -\frac{48}{65}, \; c_1 = -\frac{2}{5}, \; d_1 = \frac{18}{5}, \; c_2 = d_2 = 0,$$

and so

$$u(r, \theta) = \frac{\ln r - \ln 3}{\ln(2/3)} + \frac{3}{5}\left(r - \frac{4}{r}\right)\cos\theta + \frac{3}{65}\left(r^2 - \frac{16}{r^2}\right)\cos 2\theta - \frac{2}{5}\left(r - \frac{9}{r}\right)\sin\theta,$$

and the solution is complete. ■

The next example is of a different type again, in that it involves the solution of Laplace's equation in a region that is *unbounded* in one direction.

EXAMPLE 18.22

Find the steady state temperature distribution $T(x, y)$ in the uniform slab of metal shown in Fig. 18.33, given that no heat sources are present in the slab and the temperatures on the boundaries are

$$T(x, 0) = T(x, a) = 0 \quad \text{for } 0 < x < \infty, \text{ and } T(0, y) = f(y),$$

where $f(y)$ is a bounded function. State any additional condition that must be imposed on $T(x, y)$ for the solution to be physically possible.

Solution As the metal is uniform and there are no heat sources present, it follows that the steady state temperature must be a solution of the Laplace equation

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0.$$

The sides of the slab are parallel to the coordinate axes, and the equation is homogeneous, so we may separate variables by setting

$$T(x) = X(x)Y(y).$$

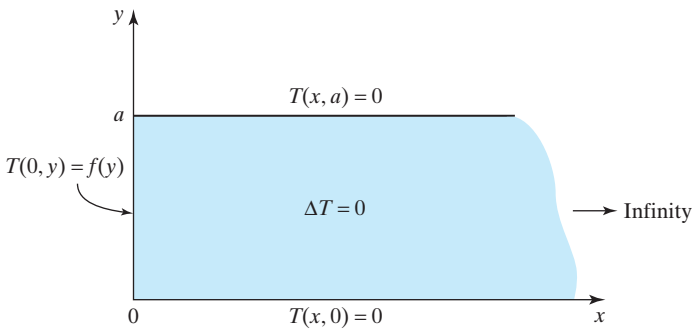


FIGURE 18.33 A semi-infinite slab of metal.

Substituting this expression into Laplace's equation and proceeding in the normal manner, we arrive at the separated form of the equation

$$\frac{Y''}{Y} = -\frac{X''}{X} = -\lambda,$$

where $\lambda > 0$ is a separation constant.

This last result separates Laplace's equation into the two differential equations

$$Y'' + \lambda Y = 0 \quad \text{and} \quad X'' - \lambda X = 0,$$

where the boundary conditions for $Y(y)$ are easily seen to be $Y(0) = Y(a) = 0$. Thus, we have arrived at the following Sturm–Liouville problem for $Y(y)$:

$$Y'' + \lambda Y = 0 \quad \text{with} \quad Y(0) = Y(a) = 0.$$

The general solution for $Y(y)$ is

$$Y(y) = A\cos\left(\sqrt{\lambda}y\right) + B\sin\left(\sqrt{\lambda}y\right).$$

Imposing these boundary conditions on the general solution for $Y(y)$ shows that the eigenvalues are $\lambda_n = n^2\pi^2/a^2$ and the corresponding eigenfunctions are $Y_n(y) = \sin(n\pi y/a)$, for $n = 1, 2, \ldots$.

Setting $\lambda = \lambda_n$ in the equation for $X(x)$ and integrating gives

$$X_n(x) = C_n \exp(-n\pi x/a) + D_n \exp(n\pi x/a).$$

To make further progress it is now necessary to recognize that when no sources are present in the metal, and a finite temperature is imposed along the boundary $x = 0, 0 < y < a$, a physically possible temperature distribution is one that must be bounded throughout the metal. This being so, we must set the coefficients $D_n = 0$ to remove the terms $\exp(n\pi x/a)$ that would otherwise become infinite as $x \to \infty$, thereby causing the functions $X_n(x)$ to simplify to $X_n(x) = \exp(-n\pi x/a)$. Notice that for convenience we have set all scale factors $C_n = 1$, since in what is to follow they will be absorbed into the new arbitrary constants d_n.

Writing $T_n(x, y) = X_n(x)Y_n(y) = \exp(-n\pi x/a)\sin(n\pi y/a)$, we now seek a solution of the form

$$T(x, y) = \sum_{n=1}^{\infty} d_n X_n(x)Y_n(y) = \sum_{n=1}^{\infty} d_n \exp(-n\pi x/a)\sin(n\pi y/a).$$

If we set $x = 0$ in this summation and use the boundary condition $T(0, y) = f(y)$, this reduces to

$$f(y) = \sum_{n=1}^{\infty} d_n \sin(n\pi y/a),$$

from which it follows in the usual manner that

$$d_n = \frac{2}{a}\int_0^a f(y)\sin\left(\frac{n\pi y}{a}\right)dy, \text{ for } n = 1, 2, \ldots.$$

The solution has been found by imposing the extra condition that $T(x, y)$ remains *bounded* in the (open) semi-infinite strip, which compensates for the normal requirement for elliptic equations that the region is closed (see page 977). ■

Other accounts of the method of separation of variables are to be found in references [3.7], [7.5], [7.7], [7.10], [7.15], [7.17], [7.19], and [7.20].

Summary

An application of the separation of variables method of solution to a PDE was seen to lead to a Sturm–Liouville problem with its parameter formed by a separation constant. When time was involved, the eigenvalues and eigenfunctions of the Sturm–Liouville problem were seen to be determined by the boundary conditions of the problem. This, in turn, was seen to determine the general structure of the solution as a series of functions of space and time, but with the multiplicative coefficients of these functions undetermined. The unknown coefficients were obtained by requiring the general series solution to satisfy the initial conditions, and by using the orthogonality properties of the functions involved. An exception was the solution of a Dirichlet problem for the Laplace equation in an annular region, where the coefficients in the series solution were obtained by matching the coefficients of corresponding sines and cosines of multiple angles. The examples given required the use of cartesian, cylindrical, and spherical polar coordinates.

EXERCISES 18.10

In Exercises 1 through 9 solve the stated boundary value problems for the wave equation in two independent variables $u_{tt} = c^2 u_{xx}$ on the interval $0 \le x \le L$.

1. A stretched string of length L, clamped at each end, starts from rest at time $t = 0$ with the initial shape $u(x, 0) = kx^2(1 - x/L)$. Find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

2. A stretched string of length L, clamped at each end, starts from rest at time $t = 0$ with the initial shape $u(x, 0) = kx(1 - x^2/L^2)$. Find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

3. A stretched string, clamped at each end, is displaced from its equilibrium position by having its mid-point given a small transverse displacement k, so that its initial shape is given by

$$u(x, 0) = \begin{cases} 2kx/L, & 0 \le x \le L/2 \\ 2k(1 - x/L), & L/2 \le x \le L. \end{cases}$$

If, while in this position, the string is released from rest at time $t = 0$, find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

4. A stretched string, clamped at each end, is displaced from its equilibrium position by having a point on the string at $x = L/3$ given a small transverse displacement k, so that its initial shape is given by

$$u(x, 0) = \begin{cases} 3kx/L, & 0 \le x \le L/3 \\ \frac{3}{2}k(1 - x/L), & L/3 \le x \le L. \end{cases}$$

If, while in this position, the string is released from rest at time $t = 0$, find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

5. A stretched string of length L, clamped at each end, starts from rest at time $t = 0$ with the initial shape $u(x, 0) = k\sin(\pi x/L)$. Use a simple argument to find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

6. At time $t = 0$ a stretched string of length L, clamped at each end, starts from its equilibrium position $u(x, 0) = 0$ with the transverse speed $u_t(x, 0) = k\sin(2\pi x/L)$. Use simple arguments to find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

7. At time $t = 0$ a stretched string of length L, clamped at both ends, starts from its equilibrium position $u(x, 0) = 0$ with the transverse speed $u_t(x, 0) = kx(1 - x/L)$. Find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

8. At time $t = 0$ a stretched string of length L, clamped at both ends, starts from its equilibrium position $u(x, 0) = 0$ with the transverse speed $u_t(x, 0) = kx^2(1 - x/L)$. Find its transverse displacement $u(x, t)$ at any subsequent time $t > 0$.

9. A string of length L is clamped at the end $x = 0$, and its other end is allowed to move along the line $x = L$ in such a way that its slope at $x = L$ remains horizontal, so that $u_x(L, t) = 0$. If the string starts from rest at the time $t = 0$ with the initial shape $u(x, 0) = kx/L$ with $0 \le x \le L$, find its transverse displacement at any subsequent time $t > 0$.

10. An approximate description of the oscillations of air caused by blowing across the end of a tube is provided by the wave equation $p_{tt} = c^2 p_{xx}$, where c is the speed of sound in air and p is the air pressure in the tube. The velocity v of the air transverse to the axis of the tube is given by $\rho v_t = -p_x$, where ρ is the density of the air. When the tube is closed at the end $x = 0$ and open at the end $x = L$, the boundary conditions are $p_x(0, t) = p_x(L, t) = 0$. Find the eigenvalues determining the possible frequencies of oscillation, the associated eigensolutions, and the transverse speed $v(x, t)$ associated with each mode.

11. Solve the initial boundary value problem $u_{xx} = u_{yy} + 5u_y$ when $u(x, 0) = e^{-6x}$ and $u_y(x, 0) = 0$. Find the approximate form of the solution when y is large and positive.

12. A rectangular membrane with its corners at $(0, 0)$, $(a, 0)$, (a, b), and $(0, b)$ has its edges clamped. Show that the eigenvalues λ_{mn} determining the vibrational frequencies $\lambda_{mn}c/2\pi$ are given by

$$\lambda_{mn}^2 = \sqrt{(n\pi/a)^2 + (m\pi/b)^2},$$

and that the corresponding eigensolutions determining the modes of vibration are proportional to

$$u_{mn}(x, y) = \sin(n\pi x/a)\sin(m\pi y/b)\cos(\lambda_{mn}ct).$$

13. The temperature $u(x, t)$ in a strip of metal of width L is governed by the heat equation $ku_{xx} = u_t$ for $0 \le x \le L$ and $t > 0$. Find the temperature in the strip given that the initial condition is $u(x, 0) = x$ and the boundary conditions, corresponding to insulated ends of the strip, are $u_x(0, t) = u_x(L, t) = 0$ for $t > 0$.

14. The electric potential $u(x, y)$ in the semi-infinite strip $x > 0, 0 < y < a$ satisfies the Laplace equation $u_{xx} + u_{yy} = 0$. Find the potential in the strip if $u(x, y)$ is finite throughout the strip and it satisfies the boundary conditions on the top and bottom of the strip

$$u_y(x, 0) = u_y(x, a) = 0,$$

corresponding to insulator sides of the strip, and the potential

$$u(0, y) = \begin{cases} 1, & 0 \le y \le a/2 \\ 0, & a/2 < y \le a \end{cases}$$

at $x = 0$ on the y-axis at the end of the strip.

15. Find the potential inside the spherical cavity in Example 18.17 when the potential on the spherical boundary $r = \rho$ is zero for $0 \le \theta < \frac{\pi}{4}$, U for $\frac{\pi}{4} < \theta < \frac{3\pi}{4}$, and zero for $\frac{3\pi}{4} < \theta \le \pi$.

16. Explain why when in spherical coordinates the solution $u(r, \theta)$ of the Laplace equation does not depend on ϕ, the solution outside a sphere on which the potential u is given can be written as a linear combination of the eigensolutions

$$u_n(r, \theta) = \frac{1}{r^{n+1}} P_n(\xi),$$

for $n = 0, 1, \ldots$, where the $P_n(\xi)$ with $\xi = \cos\theta$ are Legendre polynomials of degree n. Use this result to find the first four terms in the Fourier–Legendre expansion of the potential $u(r, \xi)$ outside a sphere of radius ρ when the potential on the surface $r = \rho$ of the sphere is zero for $0 \le \theta < \frac{\pi}{4}$, U for $\frac{\pi}{4} < \theta < \frac{\pi}{2}$, and zero for $\frac{\pi}{2} < \theta \le \pi$.

17. A uniform rectangular membrane $0 \le x \le c, 0 \le y \le d$ is clamped around its edges and performs small oscillations governed by the equation $c^2(u_{xx} + u_{yy}) = u_{tt}$, where $u(x, y, t)$ is the displacement of the membrane normal to the (x, y)-plane at time t and position (x, y), and c is a constant. Derive a general series expansion for $u(x, y, t)$ when the membrane satisfies the boundary conditions

$$u(0, y, t) = u(c, y, t) = 0 \text{ for } 0 \le y \le d$$
$$\text{and} \quad u(x, 0, t) = u(x, d, t) = 0 \quad \text{for } 0 \le x \le c$$

and the initial conditions

$$u(x, y, 0) = f(x, y) \quad \text{and} \quad u_t(x, y, 0) = g(x, y).$$

Use the result to find the form of the solution when

$$f(x, y) = 2\sin\left(\frac{3\pi x}{c}\right)\sin\left(\frac{\pi y}{d}\right) \quad \text{and } g(x, y) = 0.$$

Explain why the solution is so simple.

18. Show that the solution of $\Delta u = 0$ in the rectangle $0 \le x \le l, 0 \le y \le L$ subject to the boundary conditions $u(0, y) = u(l, y) = 0$ and $u(x, 0) = \sin(\pi x/l)$ and $u(x, L) = \sin(2\pi x/l)$ is given by

$$u(x, y) = \frac{\sinh(2\pi y/l)}{\sinh(2\pi L/l)} \sin\left(\frac{2\pi x}{l}\right) - \frac{\sinh(\pi(y - L)/l)}{\sinh(\pi L/l)} \sin\left(\frac{\pi x}{l}\right).$$

19. Show that the solution of the diffusion equation $u_t = \kappa^2 u_{xx}$ for $0 \le x \le L, t > 0$ subject to the boundary conditions

$$u(0, t) = u(L, t) = 0, \quad t > 0,$$

and the initial condition

$$u(x, 0) = \begin{cases} x, & 0 \le x \le L/2 \\ L - x, & L/2 \le x \le L \end{cases}$$

is

$$u(x, t) = \frac{4L}{\pi^2} \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^2} \times \exp\left[-\frac{(2n+1)^2 \pi^2 \kappa^2}{L^2} t\right] \sin \frac{(2n+1)\pi x}{L}.$$

20. Solve the Laplace equation

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0$$

in the annulus $3 \le r \le 5$, subject to the Dirichlet conditions

$$u(3, \theta) = 2 + \cos\theta \text{ and } u(5, \theta) = 1 - \sin 2\theta.$$

21. Find the steady state temperature distribution determined by the Laplace equation

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$$

in the semi-infinite block of metal $x \ge 0, 0 \le y \le \pi$ subject to the boundary conditions

$$T(x, 0) = T(x, \pi) = 0 \quad \text{for} \quad 0 \le x < \infty$$
$$\text{and} \quad T(0, y) = y\cos(y - \pi/2).$$

18.11 Some General Results for the Heat and Laplace Equations

(a) Equations Reducible to the Heat Equation

The simplest form of the heat equation for the function $u(x, t)$ occurs when the thermal conductivity κ is a constant and $\kappa = 1$, so the equation becomes

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}. \tag{143}$$

The following transformations reduce the given form of parabolic equation to the form given in (143).

PDEs that can be reduced to the heat equation

(i) The transformation $\tau = \kappa^2 t$ reduces the equation

$$\frac{\partial u}{\partial t} = \kappa^2 \frac{\partial^2 u}{\partial x^2} \quad \text{to} \quad \frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial x^2}.$$

(ii) The transformation $v(x, t) = \exp(-at)u(x, t)$ reduces the equation

$$\frac{\partial v}{\partial t} = \kappa^2 \frac{\partial^2 v}{\partial x^2} - ae^{at} v \quad \text{to} \quad \frac{\partial u}{\partial t} = \kappa^2 \frac{\partial^2 u}{\partial x^2}.$$

(iii) The transformation $v(x,t) = \exp[b(x - \frac{1}{2}bt)/(2\kappa^2)]u(x,t)$ reduces the equation

$$\frac{\partial v}{\partial t} = \kappa^2 \frac{\partial^2 v}{\partial x^2} - bv_x \quad \text{to} \quad \frac{\partial u}{\partial t} = \kappa^2 \frac{\partial^2 u}{\partial x^2}.$$

(iv) Successive applications of transformations (i), (ii), and (iii) reduce the equation

$$\frac{\partial w}{\partial t} = \kappa^2 \frac{\partial^2 w}{\partial x^2} - bw_x - aw \quad \text{to} \quad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

(b) The Weak Maximum/Minimum Principle for the Heat Equation

Physical intuition suggests that because heat flows from a region of high temperature to one of lower temperature, the temperature $u(x,t)$ at any interior point of the interval $0 \le x \le L$ at a time $t_0 > 0$ must be less than the maximum of the initial temperature distribution on the interval when $t = 0$, or the maximum at the ends $x = 0$ and $x = L$ during the time $0 < t < t_0$. Conversely, the temperature $u(x,t)$ in the time interval $0 < t < t_0$ will be greater than the least of the minima of the temperature distributions over the interval at the initial time, and at the ends $x = 0$ and $x = L$. These observations form the substance of Theorem 18.1, which is called the **weak maximum/minimum principle** for the heat equation. The theorem is useful when proving general properties of the heat equation, and also for finding bounds on the solution without the need to solve the equation. The proof of the theorem that follows is based on the approach used by Petrovsky.

THEOREM 18.1

The maximum/minimum principle for the heat equation Let $u(x,t)$ be the solution of the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

in the rectangular region D formed by $0 \le x \le L, 0 \le t \le t_0$, and subject to the boundary conditions

the form taken by the max/min principle for the heat equation

$$u(0,t) = h_1(t) \quad \text{and} \quad u(L,t) = h_2(t) \ \text{ for } \ 0 \le t \le t_0$$

and the initial condition

$$u(x,0) = \Phi(x).$$

Let m and M, respectively, be the smallest and greatest values assumed by u on the partial boundary Γ of the rectangle D formed by the interval $0 \le x \le L$ on the x-axis and the two vertical lines $x = 0, 0 \le t \le t_0$ and $x = L, 0 \le t \le t_0$, the line forming the top of the rectangle being omitted. Then the solution $u(x,t)$ is such that

$$m \le u(x,t) \le M.$$

Proof Let M be the maximum of $u(x,t)$ in D and Γ, and m be the minimum of u on Γ. Assume, if possible, that the statement of the theorem is false and there

exists a solution $u(x, t)$ such that $M > m$ at some point (ξ, τ) strictly inside D. Now consider the function

$$v(x, t) = u(x, t) + \frac{M - m}{4L^2}(x - \xi)^2.$$

Then on Γ we have

$$v(\xi, \tau) \leq m + \frac{1}{4}(M - m) < \frac{1}{4}M + \frac{3}{4}m = kM,$$

where $0 < k < 1$ and $v(\xi, \tau) = M$.

This shows that v does not assume its maximum value on Γ, so it must occur at some point (ξ_1, τ_1) inside D. From the elementary calculus of maxima of twice continuously differentiable functions of two variables, we must have $\partial^2 v/\partial x^2 \leq 0$ and $\partial v/\partial t \geq 0$ at (ξ_1, τ_1). Consequently, at the point (ξ_1, τ_1) we have shown that

$$\frac{\partial v}{\partial t} - \frac{\partial^2 v}{\partial x^2} \geq 0,$$

but direct calculation shows that

$$\frac{\partial v}{\partial t} - \frac{\partial^2 v}{\partial x^2} = -\frac{M - m}{2L^2} < 0.$$

This is a contradiction, so the assumption that the maximum of $u(x, t)$ can occur inside D is false. The result concerning the minimum of $u(x, t)$ follows by applying the preceding result to $-u(x, t)$, so the theorem is proved. ■

An almost immediate consequence to this theorem is the continuous dependence of the solution of the heat equation on the boundary and initial conditions, showing that it is a properly posed problem.

THEOREM 18.2

showing the continuous dependence of the solution of the heat equation on the initial and boundary conditions

The continuous dependence of $u(x, t)$ on the boundary and initial conditions Consider the two problems

(I) $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

in the rectangular region D formed by $0 \leq x \leq L, 0 \leq t \leq t_0$, and subject to the boundary conditions

$$u(0, t) = h_1(t) \quad \text{and} \quad u(L, t) = h_2(t) \text{ for } 0 \leq t \leq t_0$$

and the initial condition

$$u(x, 0) = \Phi(x),$$

and

(II) $$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}$$

in the rectangular region D formed by $0 \leq x \leq L, 0 \leq t \leq t_0$, and subject to the boundary conditions

$$v(0, t) = H_1(t) \quad \text{and} \quad v(L, t) = H_2(t) \text{ for } 0 \leq t \leq t_0$$

and the initial condition

$$v(x, 0) = \Omega(x).$$

Then, if for some arbitrarily small number $\varepsilon > 0$

$$|h_1(t) - H_1(t)| \le \varepsilon \quad \text{and } |h_2(t) - H_2(t)| \le \varepsilon \quad \text{for } 0 \le t \le t_0,$$

and

$$|\Phi(x) - \Omega(x)| \le \varepsilon \quad \text{for } 0 \le x \le L,$$

it follows that $|u(x,t) - v(x,t)| \le \varepsilon$ for $0 \le x \le L$ and $0 \le t \le t_0$.

Proof Set $w(x,t) = u(x,t) - v(x,t)$, and notice that as the heat equation is linear, $w(x,t)$ will also be a solution of the heat equation. It then follows from the boundary conditions that

$$|w(0,t)| = |h_1(t) - H_1(t)| \le \varepsilon \quad \text{and } |w(L,t)| = |h_2(t) - H_2(t)| \le \varepsilon \quad \text{for } 0 \le t \le t_0,$$

and from the initial conditions that

$$|w(x,0)| = |\Phi(x) - \Omega(x)| \le \varepsilon \quad \text{for } 0 \le x \le L.$$

From Theorem 18.1, the maximum of $w(x,t)$ on the partial boundary Γ defined in the theorem cannot exceed ε and it cannot be less than $-\varepsilon$, so $-\varepsilon \le w(x,t) \le \varepsilon$. This is equivalent to

$$|u(x,t) - v(x,t)| \le \varepsilon,$$

so the theorem is proved. ■

To see how Theorem 18.1 can be used to place bounds on solutions of the heat equation $u_t = u_{xx}$, consider the problem corresponding to $h_1(t) = t \sin t$ and $h_2(t) = 0$ for $0 \le t \le \frac{\pi}{2}$ and $\Phi(x) = \sin(3x/2) - \sin x$ for $0 \le x \le \pi$.

The maximum and minimum values of $h_1(t)$ for $0 \le t \le \frac{\pi}{2}$ are $\frac{\pi}{2}$ and 0, respectively, and $h_2(t)$ is identically zero, whereas on the interval $0 \le x \le \pi$ a plot of $\Phi(x)$ shows it has a maximum of 0.2233 at $x = 0.6858$ and a minimum of -1.2160 at $x = 2.7084$. The partial boundary Γ in Theorem 18.1 comprises the interval $0 \le x \le \pi$ on the x-axis, and the two vertical lines $x = 0$ and $x = \pi$ for $0 \le t \le \frac{\pi}{2}$, so from Theorem 18.1

$$-1.2160 \le u(x,t) \le \pi/2 \quad \text{for } 0 \le x \le \pi \quad \text{and} \quad 0 \le t \le \frac{\pi}{2}.$$

(c) The Fundamental Solution of the Heat Equation

It was proved in Section 10.2, using the Fourier transform, that when the heat equation defined in the infinite interval $-\infty < x < \infty$ is written in the form

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{k}\frac{\partial u}{\partial t} \qquad (k = \kappa^2), \tag{144}$$

the fundamental solution of the heat equation and the delta function

its solution subject to the initial condition $u(x,0) = f(x)$ is given by

$$u(x,t) = \sqrt{\frac{1}{4\pi kt}} \int_{-\infty}^{\infty} f(x') \exp\left\{-\frac{(x-x')^2}{4kt}\right\} dx'. \tag{145}$$

Setting $f(x) = \delta(x)$, where $\delta(x)$ is the Dirac delta function, simplifies this result to

$$u(x,t) = \sqrt{\frac{1}{4\pi kt}} \exp\left\{-\frac{x^2}{4kt}\right\}.$$

This elementary solution, which corresponds to an initial condition in the form of a single delta function located at the origin, is called the **fundamental solution** of the heat equation, and it is often denoted by $K(x,t)$, so that

$$K(x,t) = \sqrt{\frac{1}{4\pi kt}} \exp\left\{-\frac{x^2}{4kt}\right\}. \tag{146}$$

In terms of $K(x,t)$, the solution of

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{k}\frac{\partial u}{\partial t}$$

subject to the initial condition $u(x,0) = f(x)$ can be written

$$u(x,t) = \int_{-\infty}^{\infty} f(x')K(x-x',t)dx',$$

showing that $u(x,t)$ is the convolution of the initial condition $f(x)$ and $K(x,t)$.

The fundamental solution plays an important role in more advanced studies of the heat/diffusion equation (see, for example, references [7.14] and [7.20]).

(d) The Maximum/Minimum Principle for Solutions of the Laplace Equation

For the sake of completeness we restate the maximum–minimum theorem for harmonic functions (solutions of the Laplace equation) that was established in Theorem 14.17 of Chapter 14.

THEOREM 18.3

The maximum/minimum theorem for harmonic functions If the function $u(x, y)$ satisfies the Laplace equation (is harmonic)

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

in some open bounded region D and continuous on its boundary Γ, then the maximum and minimum values of u occur on Γ. ■

again the max/min theorem for harmonic functions and continuous dependence on Dirichlet conditions

An argument similar to the one used in Theorem 18.2 establishes the continuous dependence of solutions of the Laplace equation on Dirichlet conditions imposed on Γ, showing that the problem is well posed.

Summary

Substitutions were given that reduce certain types of parabolic equation to the standard heat equation. A maximum/minimum theorem was proved for the heat equation, and used to show the continuous dependence of the solution on the initial and boundary conditions. The delta function was then employed to derive the fundamental solution of the heat equation that enables the solution to be found subject to an arbitrary initial condition for a problem defined in the infinite interval $-\infty < x < \infty$.

18.12 An Introduction to Laplace and Fourier Transform Methods for PDEs

The solution of partial differential equations by means of Laplace and Fourier transforms has already been illustrated in Section 7.3(e)(ii) and Section 10.2. In the examples just mentioned, the application of the Fourier transform, the Fourier sine transform, and the Laplace transform to the one-dimensional heat equation all involved the same three fundamental steps that are typical of transform methods, so these are summarized below in terms of a function $u(x, t)$ that satisfies a linear constant coefficient PDE.

Steps in the solution of a PDE by means of an integral transform

the basic steps to be followed when solving a PDE using an integral transform

STEP 1 Let the solution of a PDE be the function $u(x, t)$ of the two independent variables x and t. Transform $u(x, t)$ with respect to one of its independent variables by means of an integral transform suited to the problem. If, for example, the transform is with respect to x, then a transformed variable $U(\alpha, t)$ is obtained, where α is the transform variable. If a Laplace transform is appropriate, the transform variable α will be s, and when a Fourier transform is appropriate, α will be ω. Rearrange the result to obtain an *ordinary* differential equation for the transformed variable $U(\alpha, t)$ where t is the single independent variable and α is a parameter.

STEP 2 Find the general solution of the ODE for $U(\alpha, t)$ as a function of t, with the transform variable α still appearing as a parameter in the solution, and use the boundary and/or initial conditions of the original problem to determine the precise form of the transform $U(\alpha, t)$.

STEP 3 Invert the transform $U(\alpha, t)$ to find the required solution $u(x, t)$. In simple cases the inversion can be performed with the help of a table of transform pairs, but in general $U(\alpha, t)$ must be inverted using the appropriate inversion integral.

The type of transform to be used, and the independent variable in $u(x, t)$ that is to be transformed, depends on the region in which the solution is required, and also on the boundary and initial conditions of the original problem. In general, the Laplace and the Fourier sine and cosine transforms can be used when the variable to be transformed is defined over the semi-infinite interval $[0, \infty)$, and a Fourier transform is used when the variable to be transformed is defined over the entire real line $(-\infty, \infty)$. If the transformed variable is defined over the semi-infinite interval $[0, \infty)$, the appropriate choice of transform is determined by the partial derivatives

that are to be transformed and the nature of the boundary and/or initial conditions of the original problem.

The following summary of the way in which derivatives transform illustrates what must be known about $u(x,t)$ in order that the necessary transforms of partial derivatives can be determined and, consequently, which transform should be used.

how partial derivatives transform when using different transforms

The transform of derivatives by different transforms

The **Laplace transforms** of $u(x,t)$ and its partial derivatives:

$$
\begin{aligned}
{}_t\mathcal{L}\{u(x,t)\} &= U(x,s) = \int_0^\infty e^{-st}u(x,t)dt \\
{}_t\mathcal{L}\left\{\frac{\partial u(x,t)}{\partial t}\right\} &= sU(x,s) - u(x,0) \\
{}_x\mathcal{L}\left\{\frac{\partial u(x,t)}{\partial x}\right\} &= sU(s,t) - u(0,t) \\
{}_t\mathcal{L}\left\{\frac{\partial^2 u(x,t)}{\partial t^2}\right\} &= s^2U(x,s) - su(x,0) - u_t(x,0) \\
{}_x\mathcal{L}\left\{\frac{\partial^2 u(x,t)}{\partial x^2}\right\} &= s^2U(s,t) - su(0,t) - u_x(0,t) \\
{}_t\mathcal{L}\left\{\frac{\partial^n u(x,t)}{\partial x^n}\right\} &= \frac{d^nU(x,s)}{dx^n}, \quad n = 1,2,\ldots.
\end{aligned}
$$

Corresponding results are easily written down for mixed and higher order derivatives using the results for the ordinary Laplace transform given in Theorem 7.3, so, for example,

$$
\begin{aligned}
{}_t\mathcal{L}\left\{\frac{\partial^2 u(x,t)}{\partial x \partial t}\right\} &= \frac{\partial}{\partial x}\int_0^\infty e^{-st}\frac{\partial u(x,t)}{\partial t}dt = \frac{\partial}{\partial x}(sU(x,s) - u(x,0)) \\
&= s\frac{dU(x,s)}{dx} - u_x(x,0).
\end{aligned}
$$

The **Fourier transform** of $u(x,t)$ and its partial derivatives:

$$
\begin{aligned}
{}_t\mathcal{F}\{u(x,t)\} &= U(x,\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty u(x,t)\exp\{-i\omega t\}dt \\
{}_x\mathcal{F}\{u(x,t)\} &= U(\omega,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty u(x,t)\exp\{-i\omega x\}dx.
\end{aligned}
$$

Here the replacement of an independent variable by ω in the transformed function U indicates that the Fourier transform has been performed with respect to that variable:

$$
\begin{aligned}
{}_t\mathcal{F}\left\{\frac{\partial^n u(x,t)}{\partial t^n}\right\} &= (i\omega)^nU(x,\omega), \quad n = 1,2,\ldots \\
{}_x\mathcal{F}\left\{\frac{\partial^n u(x,t)}{\partial t^n}\right\} &= \frac{\partial^n U(\omega,t)}{\partial t^n}, \quad n = 1,2,\ldots \\
{}_t\mathcal{F}\left\{\frac{\partial^n u(x,t)}{\partial x^n}\right\} &= \frac{\partial^n U(x,\omega)}{\partial x^n}, \quad n = 1,2,\ldots.
\end{aligned}
$$

Corresponding results apply when mixed partial derivatives are involved so, for example,

$$ {}_t\mathcal{F}\left\{\frac{\partial^2 u(x,t)}{\partial x \partial t}\right\} = \frac{\partial}{\partial x}\,{}_t\mathcal{F}\left\{\frac{\partial u(x,t)}{\partial t}\right\} = i\omega\frac{\partial U(x,\omega)}{\partial x}. $$

The **Fourier sine and cosine transforms** of $u(x,t)$ and its partial derivatives:

$$ {}_x\mathcal{F}_C\left\{\frac{\partial f(x,t)}{\partial x}\right\} = \omega F_S(\omega,t) - \sqrt{\frac{2}{\pi}}\,f(0,t) $$

$$ {}_x\mathcal{F}_S\left\{\frac{\partial f(x,t)}{\partial x}\right\} = -\omega F_C(\omega,t) $$

$$ {}_x\mathcal{F}_C\left\{\frac{\partial^2 f(x,t)}{\partial x^2}\right\} = -\omega^2 F_S(\omega,t) - \sqrt{\frac{2}{\pi}}\,f_x(0,t) $$

$$ {}_x\mathcal{F}_S\left\{\frac{\partial^2 f(x,t)}{\partial x^2}\right\} = -\omega^2 F_S(\omega,t) + \omega\sqrt{\frac{2}{\pi}}\,f(0,t) $$

$$ {}_x\mathcal{F}_C\left\{\frac{\partial^n f(x,t)}{\partial t^n}\right\} = \frac{\partial^n F_C(\omega,t)}{\partial t^n}. $$

Corresponding results can be written down for the transform of higher order partial derivatives and also when the transform is with respect to t instead of x. The transforms of mixed partial derivatives are obtained straightforwardly from the preceding results so that, for example,

$$ {}_x\mathcal{F}_C\left\{\frac{\partial^2 f(x,t)}{\partial x \partial t}\right\} = \frac{\partial}{\partial t}\,{}_x\mathcal{F}_C\left\{\frac{\partial f(x,t)}{\partial x}\right\} = \omega\frac{\partial F_S(\omega,t)}{\partial t} - \sqrt{\frac{2}{\pi}}\,f_t(0,t). $$

The examples that follow illustrate the use of different integral transforms when solving some simple but typical problems.

EXAMPLE 18.23

finding some solutions using integral transforms

Use a transform method to obtain the Poisson integral formula

$$ u(x,y) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{y f(\xi)}{(x-\xi)^2 + y^2}\,d\xi, $$

which solves the boundary value problem for the Laplace equation $u_{xx} + u_{yy} = 0$ in the half-plane $-\infty < x < \infty$, $y > 0$ subject to the boundary condition $u(x,0) = f(x)$.

Solution As x belongs to the entire real line $-\infty < x < \infty$, only the Fourier transform with respect to x can be used. Setting ${}_x\mathcal{F}\{u(x,y)\} = U(\omega,y)$ and transforming the Laplace equation with respect to x gives

$$ (i\omega)^2 U(\omega,y) + \frac{d^2}{dy^2}U(\omega,y) = 0. $$

This has the general solution

$$ U(\omega,y) = A(\omega)e^{\omega y} + B(\omega)e^{-\omega y}, $$

where $A(\omega)$ and $B(\omega)$ are functions of ω that are to be determined. As $y > 0$, and the solution must be bounded for both positive and negative ω, this can only be possible

if $A(\omega) = 0$ when $\omega > 0$ and $B(\omega) = 0$ when $\omega < 0$. Defining $C(\omega) = A(\omega) + B(\omega)$ allows the transform $U(\omega, y)$ to be written

$$U(\omega, y) = C(\omega)e^{-y|\omega|}, \text{ for } -\infty < \omega < \infty \quad \text{and } y > 0.$$

Provided $f(x)$ has a Fourier transform $\mathcal{F}\{f(x)\} = F(\omega)$, the result of transforming $u(x, 0) = f(x)$ is $U(\omega) = F(\omega)$. Setting $y = 0$ in $U(\omega, y)$ and using this last result shows that $C(\omega) = F(\omega)$, and so

$$U(\omega, y) = F(\omega)e^{-y|\omega|}.$$

The result of Example 10.3(c) can be rewritten as

$${}_x\mathcal{F}\left\{\sqrt{\frac{2}{\pi}}\left(\frac{y}{x^2 + y^2}\right)\right\} = e^{-y|\omega|},$$

so applying the convolution theorem to $U(\omega, y)$ and using the foregoing result yields the Poisson integral formula

$$u(x, y) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{yf(\xi)}{(x - \xi)^2 + y^2} d\xi. \quad \blacksquare$$

EXAMPLE 18.24 Use a transform method to derive the D'Alembert formula

$$u(x, t) = \frac{h(x - ct) + h(x + ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} k(\sigma)d\sigma,$$

which solves the initial value problem for the wave equation $u_{tt} = c^2u_{xx}$ with $u(x, 0) = h(x)$ and $u_t(x, 0) = k(x)$, where $-\infty < x < \infty, t > 0$.

Solution As x belongs to the entire real line $-\infty < x < \infty$, only the Fourier transform with respect to x can be used. Setting ${}_x\mathcal{F}\{u(x, t)\} = U(\omega, t)$ and transforming the wave equation with respect to x gives

$$\frac{d^2U(\omega, t)}{dt^2} = c^2(i\omega)^2U(\omega, t).$$

This ordinary differential equation in which ω appears as a parameter has the general solution

$$U(\omega, t) = A(\omega)\cos(\omega ct) + B(\omega)\sin(\omega ct),$$

where the functions $A(\omega)$ and $B(\omega)$ of ω are to be determined.

Provided $h(x)$ has the Fourier transform $\mathcal{F}\{h(x)\} = H(\omega)$, the result of transforming the first initial condition $u(x, 0) = h(x)$ with respect to x is

$${}_x\mathcal{F}\{u(x, 0)\} = H(\omega).$$

Differentiation of $U(\omega, t)$ with respect to t gives

$$\frac{\partial U(\omega, t)}{\partial t} = -\omega cA(\omega)\sin(\omega ct) + \omega cB(\omega)\cos(\omega ct),$$

and so

$$U_t(\omega, 0) = \omega cB(\omega).$$

Provided $k(x)$ has the Fourier transform $\mathcal{F}\{k(x)\} = K(\omega)$, as ${}_x\mathcal{F}\{u_t(x,t)\} = U_t(\omega,t)$ and $u_t(x,0) = k(x)$, we see that $U_t(\omega,0) = K(\omega)$. Using these results in the expression for $U(\omega,t)$ we find that the Fourier transform of the solution is

$$U(\omega,t) = H(\omega)\cos(\omega ct) + K(\omega)\frac{\sin(\omega ct)}{\omega c}.$$

If we replace $\cos(\omega ct)$ by $\frac{1}{2}(e^{i\omega ct} + e^{-i\omega ct})$, this becomes

$$U(\omega,t) = \frac{1}{2}H(\omega)(e^{i\omega ct} + e^{-i\omega ct}) + K(\omega)\frac{\sin(\omega ct)}{\omega c}.$$

The solution is now obtained by finding ${}_x\mathcal{F}^{-1}\{U(\omega,t)\}$. The transform $U(\omega,t)$ is sufficiently simple that the inversion of the first group of terms can be performed using Fourier transform pairs and Theorem 10.8, while the inversion of the last term can be obtained with the help of Example 10.3(a) and the convolution theorem. From Theorem 10.8(ii) the inverse transform of the first group of terms is seen to be

$${}_x\mathcal{F}^{-1}\left\{\frac{1}{2}H(\omega)(e^{i\omega ct} + e^{-i\omega ct})\right\} = \frac{1}{2}[h(x+ct) + h(x+ct)],$$

while appeal to Example 10.3(a) and the convolution theorem shows that

$${}_x\mathcal{F}^{-1}\left\{K(\omega)\frac{\sin(\omega ct)}{\omega c}\right\} = \frac{1}{2c}\int_{x-ct}^{x+ct} k(\sigma)\,d\sigma.$$

The D'Alembert formula now follows by addition of these results. ■

EXAMPLE 18.25

Use a transform method to find the solution of the modified wave equation

$$v_{xx} = c^2 v_{tt} + 2ckv_t + k^2 v$$

that remains finite for $t > 0$ and satisfies the initial conditions $v(x,0) = 0$ and $v_t(x,0) = 0$ and the boundary condition $v(0,t) = \sin t$ for $t > 0$.

Solution Although both x and t lie in semi-infinite intervals, only the initial conditions imposed on $v(x,t)$ are sufficient to allow the Laplace transform of the PDE to be taken with respect to t. Defining ${}_t\mathcal{L}\{v(x,t)\} = V(x,s)$, using the initial conditions $v(x,0) = 0$ and $v_t(x,0) = 0$, and taking the Laplace transform of the PDE with respect to t gives

$$\frac{d^2V(x,s)}{dx^2} = c^2s^2V(x,s) + 2cksV(x,s) + k^2V(x,s),$$

so

$$\frac{d^2V(x,s)}{dx^2} - (cs+k)^2V(x,s) = 0.$$

This ordinary differential equation with s appearing as a parameter has the solution

$$V(x,s) = A(s)\exp[(cs+k)x] + B(s)\exp[-(cs+k)x],$$

where the functions $A(s)$ and $B(s)$ of s are to be determined. For the solution to remain bounded for all t it is necessary that $A(s) = 0$, and so when $x = 0$

$$V(0,s) = B(s)\exp[-(cs+k)x].$$

Taking the Laplace transform of the boundary condition gives

$$_t\mathcal{L}\{v(0,t)\} = \mathcal{L}\{\sin t\} = 1/(s^2+1),$$

and so $B(s) = 1/(s^2+1)$ and

$$V(x,s) = \frac{1}{s^2+1}\exp[-(cs+k)x] = e^{-kx}\frac{e^{-cxs}}{s^2+1}.$$

Using the table of transform pairs and the second shift theorem to invert the Laplace transform $V(x,s)$, we arrive at the solution

$$v(x,t) = e^{-kx}\sin(t-cx)H(t-cx),$$

where H is the Heaviside unit step function.

Examination of the form of the solution shows it to be a traveling wave that decays exponentially with distance, and because of the delay introduced by the Heaviside unit step function, the periodic disturbance at $x = 0$ will have no effect at a position $x = x_0$ until a time t such that $t > cx_0$. ■

EXAMPLE 18.26

Use an integral transform to find the solution of the two-dimensional Laplace equation $u_{xx} + u_{yy} = 0$ in the infinite strip $0 \le y \le a$, given that $u(x,0) = 0$ and $u(x,a) = f(x)$, and interpret the result in terms of two different physical problems.

Solution As $-\infty < x < \infty$, it is necessary to use the Fourier transform with respect to x, so transforming the Laplace equation we find that

$$(i\omega)^2 U(\omega, y) + \frac{d^2U(\omega, y)}{dy^2} = 0.$$

The solution of this ODE for the Fourier transform $U(\omega, y)$ of the solution $u(x, y)$ is

$$U(\omega, y) = A(\omega)e^{\omega y} + B(\omega)e^{-\omega y},$$

where the functions $A(\omega)$ and $B(\omega)$ of ω are to be determined. Assuming that $f(x)$ has the Fourier transform $F(\omega)$, the Fourier transform of the boundary conditions becomes

$$_x\mathcal{F}\{u(x,0)\} = U(\omega,0) = 0 \quad \text{and} \quad _x\mathcal{F}\{u(x,a)\} = U(\omega,a) = F(\omega).$$

The transform $U(\omega, y)$ is required to satisfy these two-point boundary conditions, and a routine calculation shows that

$$U(\omega, y) = F(\omega)\frac{\sinh(\omega y)}{\sinh(\omega a)}.$$

Applying the Fourier inversion integral to $U(\omega, y)$ gives

$$u(x,y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} U(\omega, y)e^{i\omega x}d\omega.$$

If $G(\omega, y)$ is defined as

$$G(\omega, y) = \frac{\sinh(\omega y)}{\sinh(\omega a)},$$

we can write

$$U(\omega, y) = F(\omega)G(\omega, y),$$

and so

$$u(x, y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) G(\omega, y) e^{i\omega x} d\omega.$$

If $g(x, y) = {}_x\mathcal{F}^{-1}\{G(\omega, y)\}$, an application of the Fourier convolution theorem to the expression on the right gives

$$u(x, y) = \frac{1}{\sqrt{2\pi}} (f * g).$$

By definition

$$g(x, y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\sinh(\omega y)}{\sinh(\omega a)} e^{i\omega x} d\omega,$$

so after expansion of the factor $e^{i\omega x}$ this becomes

$$g(x, y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\sinh(\omega y)}{\sinh(\omega a)} \cos(\omega x) d\omega + \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\sinh(\omega y)}{\sinh(\omega a)} \sin(\omega x) d\omega.$$

The last integral is zero because its integrand is an odd function of ω, but the integrand of the first integral is an even function of ω, so

$$g(x, y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\sinh(\omega y)}{\sinh(\omega a)} \cos(\omega x) d\omega = \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} \frac{\sinh(\omega y)}{\sinh(\omega a)} \cos(\omega x) d\omega.$$

Using these results in the convolution theorem now gives

$$u(x, y) = \frac{1}{\sqrt{2\pi}} (f * g) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{2}{\pi}} \int_{\omega=0}^{\infty} \int_{-\infty}^{\infty} f(\tau) \frac{\sinh(\omega y)}{\sinh(\omega a)} \cos[(\omega - \tau)x] d\tau d\omega,$$

and so

$$u(x, y) = \frac{1}{\pi} \int_{\omega=0}^{\infty} \int_{-\infty}^{\infty} f(\tau) \frac{\sinh(\omega y)}{\sinh(\omega a)} \cos[(\omega - \tau)x] d\tau d\omega.$$

One physical interpretation of this problem is that it provides the steady state temperature distribution in a slab of metal of thickness a when the lower face is maintained at a temperature $u(x, 0) = 0$ and the upper face is maintained at the temperature $u(x, a) = f(x)$. Another interpretation is that it provides the potential distribution in air between two parallel conducting plates a distance a apart, when the lower plate is maintained at zero potential and the upper one is maintained at the potential $u(x, a) = f(x)$. ■

Fourier and Laplace transform methods for the solution of PDEs are also discussed in references [3.8] and [7.14].

Summary

The basic steps to be followed when attempting to solve a PDE by means of an integral transform were outlined, and the way in which partial derivatives are transformed by different integral transforms was listed. The examples that followed showed how the nature of the problem to be solved, together with the boundary and initial conditions, serves to determine the appropriate form of transform that is to be used.

EXERCISES 18.12

1. Find the solution $T(x,t)$ that is finite for all $x > 0$, $t > 0$ and such that $T_t = kT_{xx}$ subject to the conditions $T(x,0) = T_0$ for $x > 0$ and $T(0,t) = 0$ for $t > 0$.

2. Find the solution $T(x,t)$ that is finite for all $x > 0$, $t > 0$ and such that $T_t = kT_{xx}$ subject to the conditions $T(x,0) = 0$ for $x > 0$ and $T(0,t) = e^{-t}$ for $t > 0$.

3. Find the solution $T(x,t)$ that is finite for all $x > 0$, $t > 0$ and such that $T_t = kT_{xx}$ subject to the conditions $T(x,0) = T_0$ for $x > 0$ and $T(0,t) = T_0 \cos at$ for $t > 0$.

4. Use the Fourier transform to solve the problem $T_t = kT_{xx}$ subject to the condition $T(x,0) = T_0/(1+x^2)$.

5. Solve $u_{tt} = c^2 u_{xx} - ku$ for $-\infty < x < \infty$, $t > 0$ subject to the conditions

$$u(x,0) = \begin{cases} U, & |x| \le 1 \\ 0, & |x| > 1 \end{cases} \quad \text{and} \quad u_t(x,0) = 0.$$

6. Find the bounded solution of $u_t = \kappa u_{xx} + Q\delta(x)$ subject to the initial condition $u(x,0) = 0$ for $t > 0$, where $\delta(x)$ is the Dirac delta function.

7. Find the bounded solution of $u_{xx} + u_{yy} = 0$ in the upper half-plane $-\infty < x < \infty$, $y > 0$ subject to the condition that $u(x,0) = f(x)$.

8. Find the bounded solution of $u_{xx} + u_{yy} = 0$ in the strip $-\infty < x < \infty, 0 < y < a$ subject to the conditions $u(x,0) = f(x)$ and $u(x,a) = 0$.

9. It was shown in Section 10.2 that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\{i\omega x - \omega^2 \kappa t\} d\omega = \sqrt{\frac{1}{4\pi\kappa t}} \exp\left\{-\frac{x^2}{4\kappa t}\right\}.$$

By differentiating this result with respect to x, show that

$${}_x\mathcal{F}_S^{-1}\{\omega \exp(-\omega^2\kappa t)\} = \frac{x}{2\sqrt{2}(\kappa t)^{3/2}} \exp\left\{-\frac{x^2}{4\kappa t}\right\}.$$

10.* Find the Fourier sine transform with respect to x of the bounded solution of the heat equation $u_t = ku_{xx}$ defined for $x > 0, t > 0$ that is subject to the initial condition $u(x,0) = 0$ and the boundary condition $u(0,t) = u_0 e^{-t}$. Use the result of Exercise 9 to show the solution $u(x,t)$ is given by

$$u(x,t) = \frac{u_0 x}{\sqrt{4\pi k}} \int_0^t \exp\left\{-\left(\tau + \frac{x^2}{4k(t-\tau)}\right)\right\} \frac{d\tau}{(t-\tau)^{3/2}}, \quad \text{for } x > 0 \text{ and } t > 0.$$

11.* Find the Fourier transform with respect to x of the bounded solution of the heat equation $T_t = kT_{xx}$ that is defined for $-\infty < x < \infty$ and $t > 0$ and such that it satisfies the initial condition

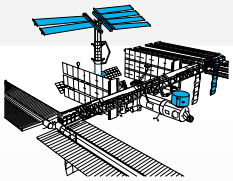
$$T(x,0) = \begin{cases} T_0, & |x| \le a \\ 0, & |x| > a. \end{cases}$$

Use result (36) of Section 10.2 to invert the Fourier transform, and express the solution in terms of the error function. Verify the solution by substituting $f(x) = T(x,0)$ in the solution for $T(x,t)$ derived in the heat conduction problem in Section 10.2.

12.* Find the Fourier transform with respect to x of the bounded solution of the heat equation $T_t = kT_{xx}$ that is defined for $-\infty < x < \infty$ and $t > 0$ and is such that it satisfies the the initial condition

$$T(x,0) = \begin{cases} T_0, & x > a \\ 0, & x < a. \end{cases}$$

Use result (36) of Section 10.2 to invert the Fourier transform, and express the solution in terms of the error function. Verify the solution by substituting $f(x) = T(x,0)$ in the solution for $T(x,t)$ derived in the heat conduction problem in Section 10.2.



CHAPTER 18
TECHNOLOGY PROJECTS

Project 1

Linear Wave Interaction

The linear wave equation $u_{tt} = c^2 u_{xx}$ with the propagation speed c has been shown to have the general solution

$$u(x,t) = f(x - ct) + g(x + ct),$$

where the functions f and g are arbitrary. The aim of this project is first to use this general solution to obtain a 3D plot showing the resolution of an initial pulse into two waves propagating in opposite directions. Then computer algebra is to be used with the general D'Alembert solution for the wave equation to make a 3D plot of the solution to a Cauchy problem with localized initial conditions.

1. Make a 3D plot showing the interaction of two waves, each with the propagation speed $c = 1$, when

$$f(x) = \begin{cases} 0, & x < -\pi/2 \\ \cos x, & -\pi/2 < x < \pi/2 \\ 0, & x > \pi/2 \end{cases}$$

$$\text{and} \quad g(x) = \begin{cases} 0, & x < -\pi/2 \\ 1, & -\pi/2 < x < \pi/2 \\ 0, & x > \pi/2. \end{cases}$$

2. Use computer algebra to find the D'Alembert solution of the wave equation $u_{tt} = u_{xx}$ when

$$u(x,0) = \begin{cases} 0, & x < -\pi/2 \\ 2\cos x, & -\pi/2 < x < \pi/2 \\ 0, & x > \pi/2 \end{cases}$$

$$\text{and} \quad u_t(x,0) = \begin{cases} 0, & x < -\pi/2 \\ x, & -\pi/2 < x < \pi/2 \\ 0, & x > \pi/2. \end{cases}$$

Make a 3D plot of the result for $-5 \le x \le 5$ and $0 \le t \le 3$ to show how the initial condition is resolved into waves propagating in opposite directions.

Project 2

Vibrating Membranes

The aim of this project is to plot the shapes of some of the eigenmodes in vibrating membranes, and to identify the nodal lines in each of these modes

1. Using the information in Example 18.12, write procedures to make 3D plots and contour plots of the eigenmodes H_{31}, H_{13}, H_{22}, and H_{23}, and in each case identify the nodal lines.

2. The eigenfunctions of a square vibrating membrane with $0 \le x \le \pi$ and $0 \le y \le \pi$ are defined by $u(m, n, x, y) = \sin(mx)\sin(ny)\cos(m^2 + n^2)$. Make a 3D plot and a contour plot of the mode u in which $m = 4$, $n = 3$, and identify the nodal lines.

Project 3

A Vibrating String Problem

The objective of this project is to write a procedure that reproduces the steps in the vibrating string problem at the start of Section 18.10, and then to make a 3D plot of the solution showing how the shape of the string changes with time.

1. Write a procedure that mimics the steps leading to the solution

$$u(x,t) = \frac{8kL^2}{\pi^3} \sum_{r=0}^{\infty} \frac{1}{(2r+1)^3} \times \sin\frac{(2r+1)\pi x}{L} \cos\frac{(2r+1)c\pi t}{L}$$

of the wave equation $u_{tt} = c^2 u_{xx}$ subject to the initial condition $u(x,0) = kx(L - x)$ and $u_t(x,0) = 0$, and the boundary conditions $u(0,t) = u(L,t) = 0$.

2. By making 3D plots of the solution with $L = \pi, c = 1$ using 5, 10, and 20 terms in the summation approximating $u(x, t)$, show that a satisfactory result is obtained by using only five terms.

Project 4

The Korteweg–de Vries Equation

The motion of long waves in shallow water is governed by the nonlinear partial differential equation

$$u_t - 6uu_x + u_{xxx} = 0,$$

called the *Korteweg–de Vries equation*, usually abbreviated to the KdV equation, where $u(x, t)$ can be considered to describe the profile of the surface wave as a function of distance x and time t. This equation, which was first derived by Korteweg and de Vries in 1895, has been shown to be of fundamental importance to various types of nonlinear wave propagation.

When the term u_{xxx} is absent from the KdV equation, it reduces to a quasilinear hyperbolic equation. It is known from Section 18.3 that the solution of a Cauchy problem for such an equation may become nonunique, and from Section 18.4 that the solution can develop into a shock wave. However, the term u_{xxx}, called a *dispersive term*, smooths the effect of the terms $u_t - 6uu_x$ in the KdV equation and balances their steepening effect and leads to the existence of smooth traveling wave solutions.

One form of smooth motion described by the KdV equation involves what is called a *solitary wave*. This is a localized disturbance in the form of the square of a hyperbolic secant function that propagates without change of shape with a speed proportional to its amplitude relative to the equilibrium water level on either side of the solitary wave. The KdV equation is first order in time, and so describes unidirectional wave propagation (propagation in one direction). Thus, if propagation is to the right, and a solitary wave of large amplitude starts well to the left of a solitary wave of smaller amplitude, the larger wave will overtake the smaller one.

The nonlinear nature of the KdV equation might be expected to cause the solution to cease to describe the propagation of such waves once interaction occurs. However, this is not the case, and after a nonlinear interaction during which the amplitudes are *not* additive, the waves reappear with their identity preserved, though with their positions slightly altered because of the interaction. This remarkable property, which occurs however many times these solitary waves interact, led to these solitary waves being called *solitons* by Zabusky and Kruskal, who were the first to observe this phenomenon as a result of numerical experiments. The interaction process is now understood analytically, but the purpose of this project is to observe this interaction and to confirm some of its qualitative features.

1. Use computer algebra to confirm by differentiation that

$$u_1(x, t) = -2\,\mathrm{sech}^2(x - 4t)$$
$$\text{and} \quad u_2(x, t) = -8\,\mathrm{sech}^2(2x - 32t)$$

are both solutions of the KdV equation $u_t - 6uu_x + u_{xxx} = 0$. Make 3D plots of the negative of $u_1(x, t)$ and $u_2(x, t)$ to show their shape and amplitude, and that their respective speeds of propagation are $dx/dt = 4$ and $dx/dt = 16$.

2. An analytical solution exhibiting soliton interaction for the KdV equation is

$$u(x, t) = -12\frac{3 + 4\cosh(2x - 8t) + \cosh(4x - 64t)}{[3\cosh(x - 28t) + \cosh(3x - 36t)]^2}.$$

Using computer algebra, substitute $u(x, t)$ into $F(x, t) = u_t - 6uu_x + u_{xxx}$, and after simplification by grouping terms show that $F(x, t) \equiv 0$, confirming that $u(x, t)$ is a solution of the KdV equation. If simplification by grouping of terms proves difficult, substitute various pairs of values of x and t into $F(x, t)$ to show that $F(x, t) = 0$, to verify that in these particular cases $u(x, t)$ is indeed a solution of the KdV equation.

3. Make a 3D plot of the negative of $u(x, t)$ for $-10 \le x \le 10$ and $-0.5 \le t \le 0.5$, using sufficient points for the plot to be relatively smooth. Choose a suitable orientation for the plot so that the crests of the propagating solitary waves are easy to follow. Notice (a) that during the interaction process around the time $t = 0$ the amplitudes are not additive, (b) that the solitons preserve their shapes after interaction, and (c) that after interaction, the path followed by the slow soliton has been slightly delayed while the path followed by the faster soliton has been slightly advanced.

4. Compare the shapes of $u_1(x, t)$ and $u_2(x, t)$ with the slow and fast solitons, respectively, both well before and after their interaction, to confirm that their shapes have been preserved.

Project 5

The Sine–Gordon Equation

This project illustrates a different type of soliton that is a solution of the nonlinear *Sine–Gordon equation*

$$u_{xx} - u_{tt} = \sin u.$$

The Sine-Gordon equation is second order in time and so describes bi-directional wave propagation (propagation in both directions).

1. Confirm by computer algebra that the function

$$u(x,t) = 4\arctan\left[\exp\left(\frac{1}{3}(5x-4t)\right)\right]$$

is a solution of the Sine–Gordon equation and, using sufficient points, make a smooth 3D plot of $u(x,t)$ for $-25 < x < 25$ and $-5 < t < 5$. This steplike function is called a *kink soliton*, and when the step changes in the opposite sense the result is called an *antikink soliton*.

2. Confirm by computer algebra that the function

$$u(x,t) = 4\arctan\left[\frac{2}{\sqrt{3}}\frac{\sinh(\sqrt{3}t)}{\cosh(2x)}\right]$$

is a solution of the Sine–Gordon equation and, using sufficient points, make a smooth 3D plot of $u(x,t)$ for $-15 < x < 15$ and $-8 < t < 8$. This shows the collision of a kink soliton and an antikink soliton.

Project 6

Dispersive Wave Propagation and the Telegraph Equation

This project demonstrates how linear equations that describe wave propagation can distort a propagating disturbance because of an effect called *dispersion*. The *telegraph equation*

$$u_{tt} - c^2u_{xx} + au_t + bu = 0,$$

with c, a, and b positive constants describes bidirectional wave propagation, and it was first derived to model telephonic communication along land lines. To see how a harmonic plane wave (a sinusoid) moving along the x-axis and governed by this equation is propagated, we consider the function $u(x,t)$ that is the real part of

$$\hat{u}(x,t) = A\exp[im(x-ct)] \quad (A \text{ real}),$$

and start by substituting $\hat{u}(x,t)$ into the telegraph equation. (This is equivalent to substituting $u(x,t) = A\cos[m(x-ct)]$ into the equation.)

Defining the *wavelength* $\lambda = 2\pi/m$, the *wave number* $k = 2\pi/\lambda$, and the frequency $\omega = 2\pi c/\lambda$ of the harmonic wave allows $\hat{u}(x,t)$ to be written

$$\hat{u}(x,t) = A\exp[i(kx-\omega t)].$$

When this expression is substituted into the telegraph equation, the following compatibility condition is found between k and ω in order that the harmonic wave is a solution of the equation:

$$\omega^2 + ia\omega - (b + c^2k^2) = 0.$$

This result is called the *dispersion relation* for the telegraph equation, and for real k it shows that ω is complex, with

$$\frac{\omega}{k} = -i\frac{a}{2k} \pm \frac{1}{2k}(4c^2k^2 + 4b - a^2)^{1/2}.$$

The quantity $kx - \omega t$ determines the *phase* of the wave, so that a wave of constant phase propagates with $kx - \omega t =$ constant, showing that the *phase velocity* of the wave is $v_P = \omega/k$. However, the dispersion relation shows that ω/k is a function of ω, so it follows that waves with different frequencies ω will propagate with different phase speeds v_p. Consequently, with the use of Fourier series, any periodic initial disturbance at time $t = 0$ can be decomposed into a sum of harmonic components, so because each component propagates with a different phase speed, when they are recombined to form the solution at later times $t_1, t_2, \ldots$, it follows that the wave shape will have changed with time. This change of shape of the wave is said to be due to *dispersion*.

When the dispersion relation is used in $\hat{u}(x,t)$, it turns out that

$$u(x,t) = \text{Re}\left\{A\exp\left(-\frac{at}{2}\right) \times \exp\left[ik\left[x \mp \frac{t}{2k}(4c^2k^2 + 4b - a^2)^{1/2}\right]\right]\right\}. \quad \text{(I)}$$

This confirms the dispersive nature of the telegraph equation, and when $a > 0$ it shows that the magnitude of the wave decays exponentially with time. If, however, $4b = a^2$ the dispersive effect vanishes and the wave propagates without change of shape, but with an exponential decay called *dissipation*. Such waves are said to be *relatively undistorted*. It was this condition that was first used to adjust the parameters in a telephone land line to remove distortion of the transmitted message due to dispersion. The decay, or dissipation, was corrected by the insertion of amplifiers at regular points along the line.

1. Let the initial wave profile be $u(x,0) = x(\pi - x)$ in the interval $0 \le x \le \pi$, and let this profile be repeated periodically along the x-axis with period π. Use computer algebra to find the coefficients $a_0, a_1, \ldots, a_6$ of the Fourier cosine series expansion of $u(x,0)$ on the interval $0 \le x \le \pi$.
2. Set $a = 0.2$, $b = 0.4$, and $c = 1$ in (I), and take the negative sign to describe a wave moving

to the right with speed $c = 1$. Let $u_k(x, t)$ denote the solution corresponding to $A = a_k$ for $k = 0, 1, \ldots, 6$, and use computer algebra to form the approximate solution of (I) given by $u_A(x, t) = \sum_{k=0}^{6} u_k(x, t)$.

3. The combined effects of dispersion and dissipation on the initial wave profile can be seen by making 2D plots of $u_A(x, t)$ at the times $t = n$ for $n = 0, 1, 2, 3$, and 4 over the respective intervals $n \le x \le n + \pi$, where the x-interval moves with speed $c = 1$ to follow the initial wave profile.
4. Repeat the calculations using $a = 0.2$, $b = 0.01$, and $c = 1$, and by again making the 2D plot in Step 3 confirm that in this case the wave decays, but is relatively undistorted (it preserves its shape as it propagates, but not its amplitude).
5. A special case of the telegraph equation is the **Klein–Gordon equation**

$$u_{tt} = au_{xx} - bu, \quad \text{with} \quad a > 0, b > 0.$$

Relate this equation to the dispersion relation in (I), and hence show that the Klein–Gordon equation is purely dispersive and so does not decay as time increases.

Project 7

Development of a Nonunique Solution

This project involves the construction of the envelope of characteristics for the first order quasilinear equation

$u_t + uu_x = 0$ subject to the initial condition $u(x, 0) = \sin x$,

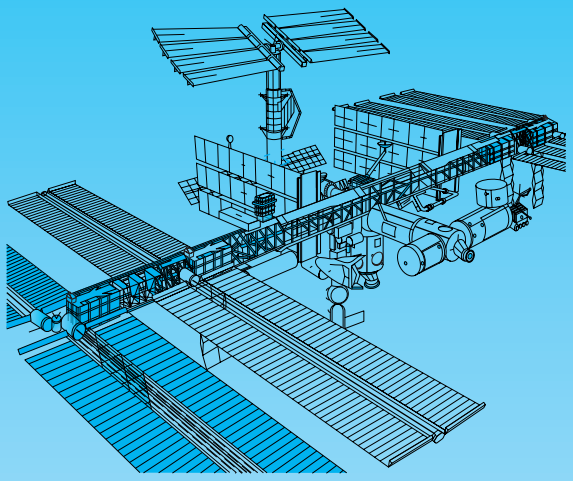
to demonstrate where and when the solution first becomes nonunique because of the intersection of characteristics. It also examines the shape of the nonlinear wave as it propagates.

1. Plot the envelope of the characteristics together with their asymptotes for the preceding problem for $0 \le x \le 2\pi$ and $0 \le t \le 4$, and confirm that its cusp forms at $x = \pi$ and $t = 1$.
2. Make 2D implicit plots of the solution $u = \sin(x - ut)$ in the interval $-5 \le x \le 5$ for the times $t = 0$, 0.5, 0.75, 1, and 2 to demonstrate how the nonuniqueness of the solution develops, using sufficient points for the plots to be smooth.
3. Make a 3D plot of the solution $u = \sin(x - ut)$ for $-2\pi \le x \le 3\pi, 0 \le t \le 3$, and $-1 \le u \le 1$ to show the global development of the nonunique solution, using sufficient points for the plot to be smooth. Compare the result with the 2D plots made in Step 2. (Hint: In the program MAPLE V, this 3D plot can be made with PDEtools and PDEplot).

This Page Intentionally Left Blank

PART EIGHT

NUMERICAL MATHEMATICS

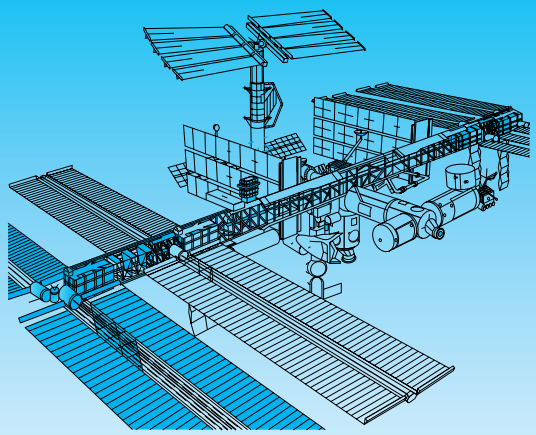


Chapter 19 Numerical Mathematics

This Page Intentionally Left Blank

CHAPTER 19

Numerical Mathematics



Unlike theoretical solutions to problems that give rise to general results that can then be related to specific problems, numerical methods only yield answers to specific problems. Because of this, numerical methods are used in the analysis of specific mathematical problems, where numerical solutions can become necessary for many different reasons. It may, for example, happen that a theoretical solution is available but is inconvenient to use, possibly because a system of linear equations arises requiring a solution that is so complex the theoretical solution is not useful. When studying a specific problem it can also happen that a definite integral occurs with no known closed form solution, or a nonlinear differential equation arises that cannot be solved theoretically. Yet another reason might be that a solution to a group of interrelated problems is so complicated that no theoretical solution is possible. In all such cases, when solving specific problems, it becomes necessary to use efficient numerical methods.

This chapter describes how to deal with the most frequently occurring types of numerical problem. These are interpolation, root finding, numerical integration, the numerical solution of large systems of linear equations, the numerical determination of eigenvalues and eigenvectors, and the numerical solution of initial value problems for linear and nonlinear differential equations and systems.

The methods described here are the classical ones, so they are neither as efficient nor as sophisticated as the methods used in currently available numerical symbolic algebra packages, though they are practical and can be used for straightforward calculations. The reason for their inclusion is because they illustrate in a concise way some of the most important general principles that are involved, while at the same time showing both the shortcomings and advantages of different methods. One essential difference between the classical methods described in this chapter, and many of the codes used in practice, is that modern codes are adaptive, so they can switch between methods of solution to speed up convergence, or adjust step size when integrating differential equations to maintain a predetermined accuracy.

19.1 Decimal Places and Significant Figures

Many of the problems that occur in engineering and physics have no analytical solution, and even when one can be found it is frequently the case that the form in which it arises is difficult to use directly if numerical results are required. There are many reasons for such limitations, some typical ones being that the zeros of a function involved in the solution cannot be found analytically, a definite integral that arises cannot be evaluated analytically, an analytical solution of a nonlinear differential equation cannot be found, or a large system of linear simultaneous equations must be solved. A situation of a different type arises when an analytical solution is known, but its application in specific cases leads to a prohibitive amount of calculation, so a more efficient numerical method becomes necessary.

As most numerical results can only be approximate, such as calculations involving $\sqrt{2}$, e, or π, it is necessary to have a simple way of indicating their accuracy. This is accomplished either by stating that a result is accurate to ***n* decimal places**, or that it is accurate to a given number of **significant digits (figures)**. For example, when approximating a number such as

decimal places

$$17.213622,$$

to *three* decimal places, the *fourth* digit after the decimal point is examined, and if the digit is 5 or more the preceding digit is increased by one and the result truncated to three places after the decimal point. However, if the *fourth* digit is 4 or less, the previous digit is left unchanged and the result is truncated to the existing three digits that follow the decimal point. When this process is applied to the above number to approximate it to an accuracy of three decimal places it becomes

$$17.214,$$

whereas if it is approximated to an accuracy of four decimal places it becomes

$$17.2136.$$

This process of approximating a number to n decimal places by increasing the nth digit by 1, if the $(n+1)$th digit is a 5 or more, and then truncating the result after n decimal places is called **rounding up** to an accuracy of n decimal places. Similarly, the process of leaving the nth digit unchanged when the $(n+1)$th digit is a 4 or less, and truncating the result after n decimal places is called **rounding down** to an accuracy of n decimal places.

rounding up and down

To express a number accurately to n significant figures involves a somewhat different argument from the one just described. The first nonzero digit that occurs in a number, irrespective of where the decimal point is located, is called the *first* (and most) *significant digit*, so in a number such as 3.496221 the first significant digit is 3, and in a number such as 0.004713 the first significant digit is 4. Starting from the first significant digit and counting $n+1$ digits to the right, the nth digit is rounded up or down, according as the $(n+1)$th digit is 5 or more, or 4 or less, as previously described. The number is then truncated after the group of n digits obtained in this way, with zeros being entered in place of any other digits that appear *before* the decimal point. This process is called expressing the number accurately to ***n* significant digits (figures)**. So, to three significant digits, the number

significant digits

$$315{,}814$$

becomes 316,000, while to four significant digits the number

$$0.004723217$$

becomes 0.004723.

Accuracy can be lost if the (approximate) result of one numerical calculation is used in a subsequent numerical calculation, and certainly if this process is repeated many times. To avoid loss of accuracy it is necessary to work to a fixed number of digits that is sufficiently large. Calculators and computers use a fixed number of digits, but symbolic algebra computer packages allow the user to choose the number so that high accuracy can be maintained throughout a sequence of calculations.

fixed and floating point numbers

The form in which numbers have been represented so far is called a **fixed point** decimal representation, because the numbers are displayed relative to the decimal point that is involved. The **floating point** representation used in most computer calculations involves writing a number x in the form

$$x = r \cdot N^s,$$

where the number N is called the **base** of the representation, the number r is called the **mantissa**, and s is called the **exponent**. The mantissa is usually chosen to have one digit in front of the decimal point. So, to the base 10, the number 453.7 has the floating point representation 4.537×10^2, while the number 0.000369 has the representation 3.69×10^{-4}. A notation used for floating point representations in machine computation to the base 10 involves representing x in floating point form by writing the mantissa r first, then the symbol E followed by the exponent s, which may be positive or negative. Most computers normalize so that the mantissa is between 0 and 1, so when using this convention the number 453.7 becomes 0.4537E3, and the number 0.000369 becomes 0.369E–3.

Summary

Accuracy in terms of decimal places and significant figures was defined, and the convention for rounding numbers up or down was explained. Floating point calculations were introduced, and the importance of expressing accuracy in terms of significant digits when working with floating point numbers was stressed.

19.2 Roots of Nonlinear Functions

Let $f(x)$ be a real valued function defined for $a \leq x \leq b$. A number ξ is called a **root** of the function $f(x)$ in this interval if $f(\xi) = 0$ and, correspondingly, a number $x = \xi$ that makes $f(x)$ vanish is called a **zero** of $f(x)$. The need to find roots of functions is fundamental to the development and application of mathematics, and only in simple cases can the roots be determined analytically, so in all other cases it is necessary to find them numerically. Many different methods exist for the numerical determination of roots of functions, but of these only the *bisection method*, the *fixed point method*, and *Newton's method* will be described in any detail, as they are in everyday use and are easily implemented on a computer.

(a) The Bisection Method

Apart from graphing $f(x)$ and finding by inspection those values of x for which $f(x) = 0$, the simplest systematic method for finding the roots of a function $f(x)$

is the **bisection method**. The method is easily programmed, and it applies to roots of functions $f(x)$ with the property that $f(x)$ changes sign when x crosses a root. The determination of a root accurately by this method depends on the ability to evaluate the function with sufficient accuracy that its sign change can be determined correctly.

To understand how the method works, consider a continuous function $f(x)$ and numbers $\alpha < \beta$ such that $f(\alpha)$ and $f(\beta)$ have opposite signs. Then from the intermediate value theorem the function $f(x)$ must vanish at least once (have at least one root) ξ between α and β, as shown in Figs. 19.1a,b. However, if $f(\alpha)$ and $f(\beta)$ have the same sign, nothing can be deduced about the existence of roots in the interval, as can be seen from Figs. 19.1c–e, which illustrate situations in which

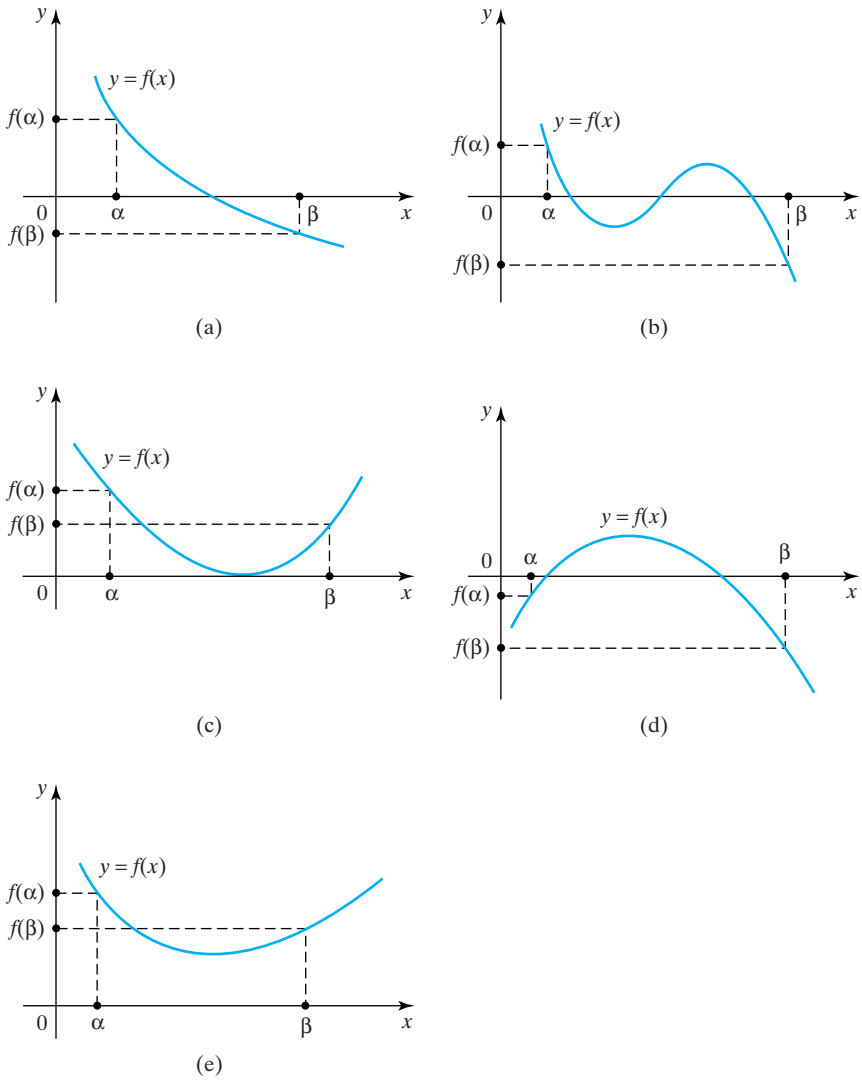


FIGURE 19.1 Roots and the product $f(\alpha)f(\beta)$ in the interval $\alpha \le x \le \beta$. (a) $f(\alpha)f(\beta) < 0$, one root. (b) $f(\alpha)f(\beta) < 0$, three roots. (c) $f(\alpha)f(\beta) > 0$, double root. (d) $f(\alpha)f(\beta) > 0$, two roots. (e) $f(\alpha)f(\beta) > 0$, no roots.

there are a double root, two roots, and no root, respectively. In what follows we will assume that $f(x)$ experiences a change of sign across the interval, and that α and β are chosen sufficiently close that there is only *one* root in the interval, as illustrated in Fig. 19.1a. When $f(x)$ is sufficiently simple this can usually be achieved by graphing $f(x)$ and selecting suitable values for α and β.

geometrical interpretation of the bisection method

To implement the bisection method, a simple test is needed to see if a function $f(x)$ has opposite signs at the ends of an interval $\alpha < x < \beta$. Such a test is provided by examining the product $f(\alpha)\,f(\beta)$, because when this is negative a sign change occurs, but when it is positive there is no such sign change. When, as may happen during a computation, a computer finds that $f(\alpha)\,f(\beta) = 0$, the value of $f(\alpha)$ must be examined to avoid interpreting as a true zero an approximate number α that causes the computer arithmetic system to regard this product function as zero.

The first step in the bisection method involves dividing (bisecting) the interval $\alpha \le x \le \beta$ into the two subintervals $\alpha < x < x_1$ and $x_1 < x < \beta$, where $x_1 = \frac{1}{2}(\alpha + \beta)$. The subinterval to be considered next is obtained by replacing α by x_1 if $f(\alpha)\,f(x_1) > 0$, because in this case $f(x)$ changes sign in the subinterval $x_1 < x < \beta$ so this interval must contain a root of $f(x)$. Conversely, if $f(\alpha)\,f(x_1) < 0$, the subinterval to be considered is obtained by replacing β by x_1, because in this case $f(x)$ experiences a change of sign in the subinterval $\alpha < x < x_1$, and so this interval must contain a root ξ. The task of finding the root has now been refined from considering the interval $\alpha \le x \le \beta$ and replaced by the task of finding the root in an interval half the size.

The bisection process involves a repetition of this procedure, each time using the smaller subinterval found at the previous stage of the calculation, so that after m steps the root ξ will be contained in an interval of length $|\alpha - \beta|/2^m$. If the root is required to be accurate to within an error of ε, where $\varepsilon > 0$ is a preassigned small quantity, machine computation that works with a fixed number of digits proceeds until the first time successive iterates x_m and x_{m+1} satisfy the condition $|x_m - x_{m+1}| < \varepsilon$. The required approximation to the root ξ is then taken to be $x_m \pm \varepsilon$.

The bisection method has the property that the bound placed on the error involved is halved at each iteration. Unlike some other methods, provided the bisection method is applicable it *always* converges to a root, though if more than one root occurs in the initial interval $\alpha \le x \le \beta$ it is not known in advance to which root the method will converge.

The bisection method has the advantage of being simple and using the minimum amount of information, because it only depends on the functional values of $f(x)$ at the end points of an interval and not on the calculation of derivatives, though other methods may converge faster. The practical implementation of the method on a computer suffers from the fact that when the product $f(\alpha)\,f(\beta)$ is determined, underflow of this floating point number becomes inevitable as the upper and lower bounds approach the root. However, this is easily overcome by determining the sign of $f(\alpha)\,f(\beta)$ by examining the signs of $f(\alpha)$ and $f(\beta)$. Because the bisection method is affected less by limiting precision, a different and faster method is often used to start the calculation, and a switch is made to the bisection method once a very accurate approximation to the root has been obtained.

The bisection method cannot be used to find a root $x = \xi$ of a function that is either convex or concave at $x = \xi$, as illustrated in Fig. 19.1e, because such functions do not change sign as x crosses ξ. This can happen, for example, when seeking the roots of polynomials of even order, the simplest case of which is $f(x) = (x - a)^2$ with a double root at $x = a$.

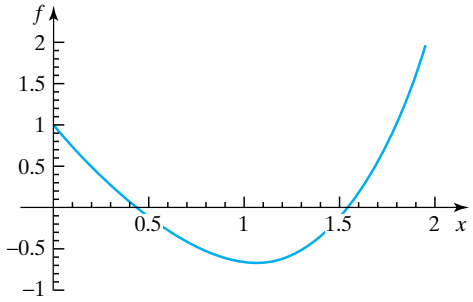


FIGURE 19.2 The function $f(x) = 1 - 3x + \frac{1}{2}xe^x$.

The numerical determination of multiple (repeated) roots is difficult, so only an outline of a possible approach will be given here for a polynomial of degree n with real roots, one of which is a double root. The difficulty that arises when seeking multiple roots is because the calculation always leads to an ill-conditioned problem—that is, to a problem in which an extremely small error in part of the calculation leads to a very large error in the result.

deflation of a polynomial

The approach we now describe involves what is called the **deflation** of the polynomial. First a single root of the polynomial is found, and the polynomial is then divided by the corresponding factor to obtain a polynomial of degree $n - 1$. A repetition of this process involving each of the $n - 2$ single roots will lead to a quadratic whose double root can then be found from the quadratic formula. When it is necessary, deflation must always be carried out with care to avoid the compounding of errors.

It is important to remember that the bisection method cannot be used to compute roots of even order, because in such cases no sign change is involved, but it works well for roots of odd order irrespective of their multiplicity. One approach to the multiple root problem involves using the bisection method with different starting intervals, and another involves using other methods with different guesses.

EXAMPLE 19.1 Use the bisection method to find the smallest root of the function $f(x) = 1 - 3x + \frac{1}{2}xe^x$.

Solution Examination of Fig. 19.2 shows that an approximation to the smallest root of $f(x) = 0$ is $x = 0.45$, and that suitable values for α and β are $\alpha = 0.43$ and $\beta = 0.47$, because $f(\alpha) = 0.0405$ and $f(\beta) = -0.0340$, and the graph shows that there is only one root between α and β.

If at each stage of the calculation the left end point of an interval containing the root ξ is denoted by x_l and the right end point by x_r, the calculation can be arranged as follows.

n	Left End Point x_l	Right End Point x_r	x_n	$f(x_l)$	$f(x_n)$	$f(x_l)\,f(x_n)$	New Interval	Approximate Root
1	$\alpha = 0.43$	$\beta = 0.47$	0.45	0.0405	0.0029	>0	$0.45 < \xi < 0.47$	0.45
2	0.45	0.47	0.46	0.0029	−0.0157	<0	$0.45 < \xi < 0.46$	0.46
3	0.45	0.46	0.455	0.0029	−0.0064	<0	$0.45 < \xi < 0.455$	0.455
4	0.45	0.455	0.4525	0.0029	−0.0018	<0	$0.45 < \xi < 0.4525$	0.4525

Continuing this process shows that to an accuracy of five decimal places the required value of the root is $x = 0.45154$. ■

(b) Fixed Point Iteration

This method is well suited to machine computation provided numerical values of the function involved are easily calculated, and a good approximation to the root is used to start the iteration process. The idea is straightforward, and its success depends on rewriting the given function $f(x)$ whose root is required in the form

$$f(x) = x - g(x). \tag{1}$$

Then if $x = \xi$ makes the expression on the right of (1) vanish, it follows that ξ is a root of $f(x)$. The representation of $f(x)$ in the form given in (1) is not unique, because as will be seen in the examples that follow, $g(x)$ can be written in more than one way. Later we will derive a simple condition on the form of $g(x)$ that must be satisfied together with the value $x_0 = \alpha$ used to start the iteration process in order that the calculations are likely to converge to the root ξ.

If we now consider the function $g(x)$ to *map* a point x into a point $g(x)$, then a root $x = \xi$ of equation (1) has the property that $g(x)$ maps the point ξ into itself, and for this reason ξ is called a **fixed point** of the equation

$$x = g(x). \tag{2}$$

fixed points and iteration

The fixed point iterative scheme follows from (2) by writing it as

$$x_{n+1} = g(x_n), \tag{3}$$

and starting the iteration process by setting $x_0 = \alpha$. The iteration will be said to **converge** if the sequence of iterates x_n approaches a limit as $n \to \infty$, and to **diverge** if no such limit exists. Suppose that when the iterations converge the result is required to be accurate to within an error of ε, where $\varepsilon > 0$ is a preassigned small quantity. Then the calculation proceeds until the first time successive iterates x_m and x_{m+1} satisfy the condition $|x_m - x_{m+1}| < \varepsilon$. The required approximation to the root ξ is then taken to be $x_n \pm \varepsilon$.

EXAMPLE 19.2

Find a fixed point iterative scheme for determining $\sqrt{a}$ when $a > 0$, and use it to calculate $\sqrt{2}$ to an accuracy of six decimal places.

Solution The required number $\sqrt{a}$ is a solution of the equation $x^2 = a$, so to express this in the form given in (2) we write it as $2x^2 = x^2 + a$, and then divide the result by $2x$ to arrive at the result

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right),$$

so in the notation of (2) the function $g(x) = \frac{1}{2}(x + \frac{a}{x})$.

The fixed point iterative scheme follows from this, as in (2), by replacing x on the left by x_{n+1} and x on the right by x_n to obtain

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

The iteration is started by setting $n = 0$ and $x_0 = k$, where k is an approximation to $\sqrt{a}$.

To illustrate the scheme we will calculate $\sqrt{2}$, so as $a = 2$ the scheme becomes

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right),$$

and for simplicity we start by setting $x_0 = 1$. The results of the calculation are

$$\begin{aligned} x_0 &= 1 \\ x_1 &= 1.5 \\ x_2 &= 1.41666667 \\ x_3 &= 1.41421569 \\ x_4 &= 1.41421356 \\ x_5 &= 1.41421356. \end{aligned}$$

As the x_4 and x_5 iterates are identical, rounding the result of x_5 to six decimal places gives $\sqrt{2} = 1.414214$. ■

The fixed point iterative scheme in Example 19.2 converged rapidly, and it is this scheme that is used in computers to determine the square root of any positive number to an accuracy that is within the capability of the computing system and software being used. Experimentation will show that this iterative scheme is stable with respect to the choice of the starting approximation, because it will always converge to $\sqrt{2}$, though a starting approximation close to $\sqrt{2}$ will, of course, lead to the most rapid convergence.

To examine iterative schemes a little further, and to show that convergence does not always occur, we consider the next example.

EXAMPLE 19.3

examination of two fixed point iterative schemes

Devise fixed point iterative schemes to find the roots of the quadratic equation

$$2x^2 - 24x + 41 = 0,$$

and test them numerically.

Solution Two obvious fixed point iterative schemes that can be obtained directly from the equation follow by first writing it in either of the forms

$$x = \frac{1}{24}(2x^2 + 41) \quad \text{or} \quad x = 12 - \frac{41}{2x}.$$

Replacing x by x_{n+1} on the left and by x_n on the right we obtain the following two schemes:

$$\text{Scheme A:} \quad x_{n+1} = \frac{1}{24}\left(2x_n^2 + 41\right), \quad \text{and} \quad \text{Scheme B:} \quad x_{n+1} = 12 - \frac{41}{2x_n}.$$

An application of the quadratic formula shows the two roots to be $x = 6 - \frac{1}{2}\sqrt{62} = 2.0630$ and $x = 6 + \frac{1}{2}\sqrt{62} = 9.9370$, so starting approximations close

to these values are $x_0 = 2$ and $x_0 = 10$. Scheme A leads to the results

$$\begin{aligned} x_0 &= 2 \\ x_1 &= 2.0417 \\ x_2 &= 2.0557 \\ x_3 &= 2.0605 \\ x_4 &= 2.0621 \\ x_5 &= 2.0627 \\ x_6 &= 2.0630 \\ &\vdots \\ x_\infty &= 2.0630. \end{aligned} \qquad \begin{aligned} x_0 &= 10 \\ x_1 &= 10.0417 \\ x_2 &= 10.1113 \\ &\vdots \\ x_8 &= 12.4801 \\ x_9 &= 14.6877 \\ &\vdots \\ x_\infty &= \infty \end{aligned}$$

Clearly Scheme A is only partially successful, because although when started with $x_0 = 2$ it converges to the zero close to 2, it diverges when started with $x_0 = 10$.

Scheme B produces the following results:

$$\begin{aligned} x_0 &= 2 \\ x_1 &= 1.75 \\ x_2 &= 0.2857 \\ x_3 &= -59.7500 \\ x_4 &= 12.3431 \\ x_5 &= 10.3392 \\ x_6 &= 10.0172 \\ x_7 &= 9.9535 \\ &\vdots \\ x_\infty &= 9.9370. \end{aligned} \qquad \begin{aligned} x_0 &= 10 \\ x_1 &= 9.7222 \\ x_2 &= 9.8914 \\ x_3 &= 9.9275 \\ x_4 &= 9.9350 \\ x_5 &= 9.9370 \\ x_6 &= 9.9370 \\ &\vdots \\ x_\infty &= 9.9370 \end{aligned}$$

Here also scheme B is also only partially successful, though this time for a different reason. Although, as required, the iterates converge to the root close to 10 when started with $x_0 = 9$, when started with $x_0 = 2$ they fail to converge to the root close to 2 and again converge to the root close to 10. ■

To understand this behavior of iterative schemes we need the following theorem that gives conditions for the choice of $g(x)$ and the starting approximation x_0 that will ensure the convergence of the scheme.

THEOREM 19.1

Convergence of a fixed point iterative scheme Let $g(x)$ be defined in the interval $a \le x \le b$ in which it has a fixed point ξ, and let $g(x)$ be continuous throughout this interval with a continuous derivative $g'(x)$ such that $|g'(x)| \le k < 1$. Then the equation $x = g(x)$ has a unique fixed point ξ in the interval, and if x_0 is such that $a \le x_0 \le b$ the iterative scheme

condition for convergence of a fixed point iterative scheme

$$x_{n+1} = g(x_n)$$

will converge to ξ.

Proof The proof involves two steps, in the first of which a fixed point ξ is assumed and shown to be unique, whereas in the second we go on to prove the convergence of the scheme and to justify the assumption of the existence of a fixed point. To show that the fixed point is unique let us assume, if possible, that two *different* fixed points ξ_1 and ξ_2 occur inside the interval, so that $\xi_1 = g(\xi_1)$ and $\xi_2 = g(\xi_2)$. Considering the expression $|\xi_1 - \xi_2|$, applying the mean value theorem, and using the condition $|g'(x)| \leq x_0 < 1$, we find that for some number η inside the interval $a \leq x \leq b$

$$|\xi_1 - \xi_2| = |g(\xi_1) - g(\xi_2)| = |g'(\eta)(\xi_1 - \xi_2)| \leq x_0|\xi_1 - \xi_2| < |\xi_1 - \xi_2|,$$

but this is impossible, so the contradiction implies the uniqueness of the fixed point.

Next, to prove the convergence of the scheme, we again make use of the mean value theorem that asserts there is some point ζ_n between x_{n-1} and ξ such that

$$|\xi - x_n| = |g(\xi) - g(x_{n-1})| = |g'(\zeta_n)(\xi - x_{n-1})| = |g'(\zeta_n)||\xi - x_{n-1}| \leq x_0|\xi - x_{n-1}|.$$

Repeated application of this inequality leads to the result $|\xi - x_n| \leq x_0^n|\xi - x_0|$, but as $0 \leq x_0 < 1$ we have $\lim_{n\to\infty} x_0^n = 0$, so that

$$\lim_{n\to\infty} |\xi - x_n| = 0, \quad \text{and hence} \quad \lim_{n\to\infty} x_n = \xi.$$

With a little more trouble, the iterates can be shown to form a Cauchy sequence, and an appeal to the completeness of real numbers then guarantees that the sequence has a limit ξ, so the theorem is proved. ■

This theorem explains the results of Example 19.2. In Scheme A the function $g(x) = \frac{1}{24}(2x^2 + 41)$, so $|g'(x)| = \frac{1}{6}|x|$ and $|g'(x)| < 1$ when $0 < x < 6$, showing the scheme to be convergent to the root close to 2 when an initial approximation close to 2 is used. However, when $x = 10$ the conditions of the theorem are not satisfied so the scheme cannot be expected to converge to the root close to 10, though it does not assert that it will diverge.

In the case of Scheme B we have $g(x) = 12 - \frac{41}{2x}$ so that $|g'(x)| = \frac{41}{2x^2}$. This shows that the scheme will converge to the root close to 10 for an x_0 close to 10, because then $|g'(x)| < 1$, but that it cannot be expected to converge to the root close to 2 where the condition is violated, though again the theorem does not assert that in this case it will diverge. It is possible to show that if $|g'(\xi)| > 1$, the iteration will not converge, except by accident.

convergent and divergent iterations

The reason for the convergence or divergence of iterative schemes is most easily understood by using a graphical representation of a fixed point iteration process. Typical cases are illustrated in Fig. 19.3, where diagrams (a) and (b) show how the mapping $x_{n+1} = g(x_n)$, using the lines $y = x$ and $y = g(x)$, can lead to *convergent* processes, while diagrams (c) and (d) show how *divergent* processes can arise.

(c) Newton's Method

Our starting point for the derivation of **Newton's method** for the determination of a zero of a differentiable function $f(x)$, also known as the **Newton–Raphson** method, is the mean value theorem representation of $f(x)$ about a point $x = x_0$ that can be written

$$f(x) = f(x_0) + (x - x_0)\, f'(\xi), \tag{4}$$

where ξ is a point between x_0 and x.

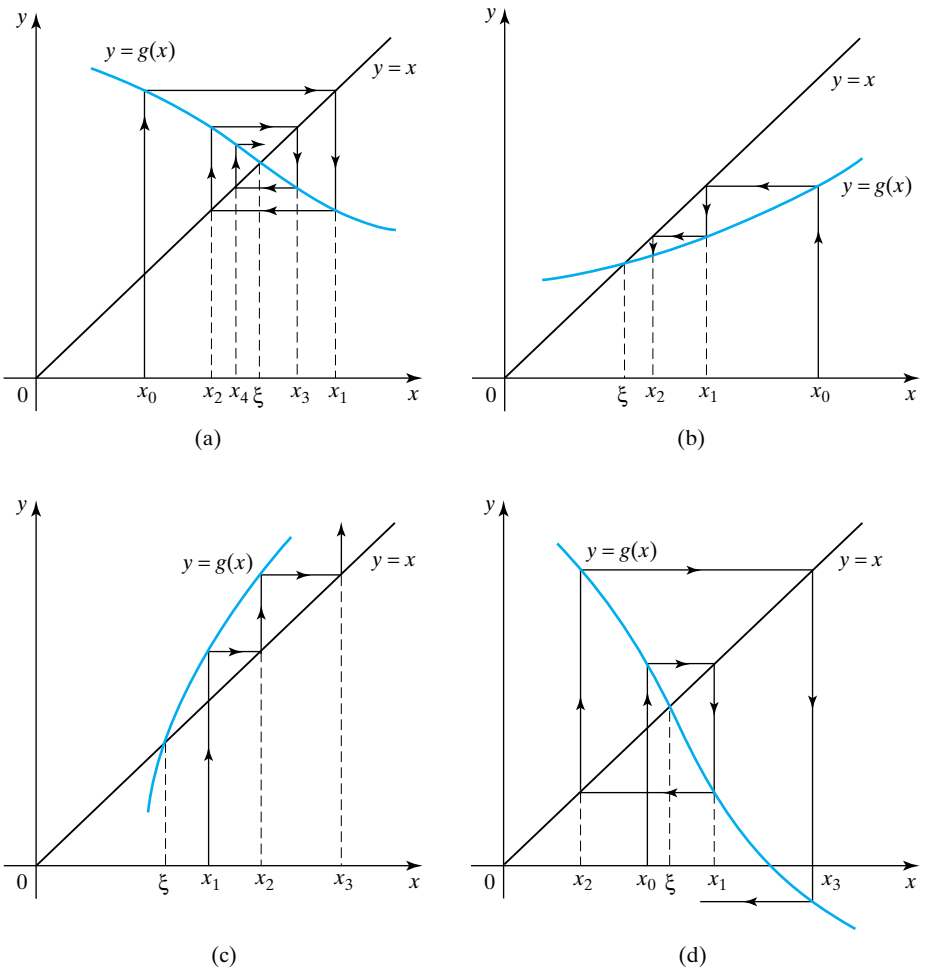


FIGURE 19.3 Typical convergent iterative processes in (a) and (b), and typical divergent iterative processes in (c) and (d).

If we set $h_0 = x - x_0$, and choose h_0 so that $x_0 + h_0$ is a zero of $f(x)$, result (4) becomes

$$h_0 = -\frac{f(\xi)}{f'(\xi)},$$

so the zero $x = x_0 + h_0$ of $f(x)$ is given by

$$x = x_0 - f(\xi)/f'(\xi). \tag{5}$$

As ξ is unknown, replacing it by x_0 produces the approximation x_1 given by

$$x_1 = x_0 - f(x_0)/f'(x_0).$$

Newton's method

Iterating this result leads to the **Newton's method**

$$x_{n+1} = x_n - f(x_n)/f'(x_n), \tag{6}$$

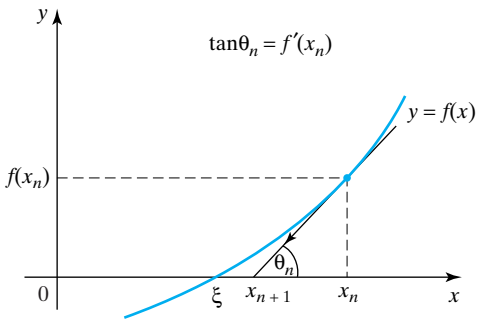


FIGURE 19.4 The tangent approximation used in Newton's method.

with $n = 0, 1, 2, \ldots$. If a tolerance ε is set, where $\varepsilon > 0$ is a preassigned small quantity, the calculations proceed until the first time the successive iterates x_m and x_{m+1} satisfy the condition $|x_m - x_{m+1}| < \varepsilon$. The number $x_{m+1} \pm \varepsilon$ is then taken to be the required approximation to the root ξ. Notice that Newton's method is a special example of fixed point iteration with $g(x) = x - f(x)/f'(x)$ and, in connection with Theorem 19.1, that the expression $|\xi - x_n| = |g'(\zeta_n)||\xi - x_{n-1}|$ tells us that $|\xi - x_n|$ approximates $|g'(\xi)||\xi - x_{n-1}|$ as the iterations converge. Clearly, the smaller $|g'(\xi)|$, the faster the convergence. For Newton's method and a simple root, this quantity is zero. So the argument suggests that for Newton's method the iterations converge faster than linearly, as is indeed the case. Typically, both fixed point iteration and Newton's method converge to the root nearest to the initial guess, though as has already been remarked, this is not true of the bisection method. Newton's method is generally much faster than the bisection method for simple roots, though not for multiple roots.

how Newton's method uses the tangent line approximation

The geometrical interpretation of Newton's method is illustrated in Fig. 19.4, where the $(n+1)$th approximation x_{n+1} is obtained from the nth approximation x_n by tracing back the tangent to the curve $y = f(x)$ at the point $(x_n, f(x_n))$ to the point x_{n+1} where it intersects the x-axis.

EXAMPLE 19.4

Use Newton's method to find the zeros of $f(x) = 1 - 3x + \frac{1}{2}xe^x$ accurate to five decimal places.

Solution A graph of $f(x)$ shows that it has zeros close to 0.5 and 1.6, so we will use these as our starting approximations. As $f'(x) = \frac{1}{2}(1+x)e^x - 3$, Newton's method becomes

$$x_{n+1} = x_n - \left(1 - 3x_n + \frac{1}{2}x_n e^{x_n}\right) \Big/ \left(\frac{1}{2}(1+x_n)e^{x_n} - 3\right) \quad \text{for } n = 0, 1, 2, \ldots.$$

Starting the calculation with $x_0 = 0.5$ gives

$$\begin{aligned} x_0 &= 0.5 & x_3 &= 0.451542 \\ x_1 &= 0.450200 & x_4 &= 0.451542 \\ x_2 &= 0.451541, \end{aligned}$$

so to an accuracy of five decimal places the smallest zero of $f(x)$ is 0.45154.

Similarly, when the calculation is started with $x_0 = 1.6$, we find that

$$\begin{aligned} x_0 &= 1.6 & x_3 &= 1.549538 \\ x_1 &= 1.552769 & x_4 &= 1.549538 \\ x_2 &= 1.549552 \end{aligned}$$

so to an accuracy of five decimal places the largest zero of $f(x)$ is 1.54954. ■

divergent and repeated cycle Newton iterations

This example illustrates the speed with which Newton's method can converge to a zero when a good starting approximation is used and the tangent to the graph $y = f(x)$ at a zero is not inclined at a small angle to the x-axis making high accuracy difficult to obtain. A poor starting approximation can cause Newton's method to diverge from the required zero, as illustrated in Fig. 19.5a where successive approximations move further away from the zero. Sometimes an unfortunate choice of starting approximation can lead to the situation illustrated in Fig. 19.5b where the iteration cycles indefinitely. To avoid situations like these, machine computations place a limit on the number of iterations to be performed to achieve the required accuracy, after which a new starting approximation must be used.

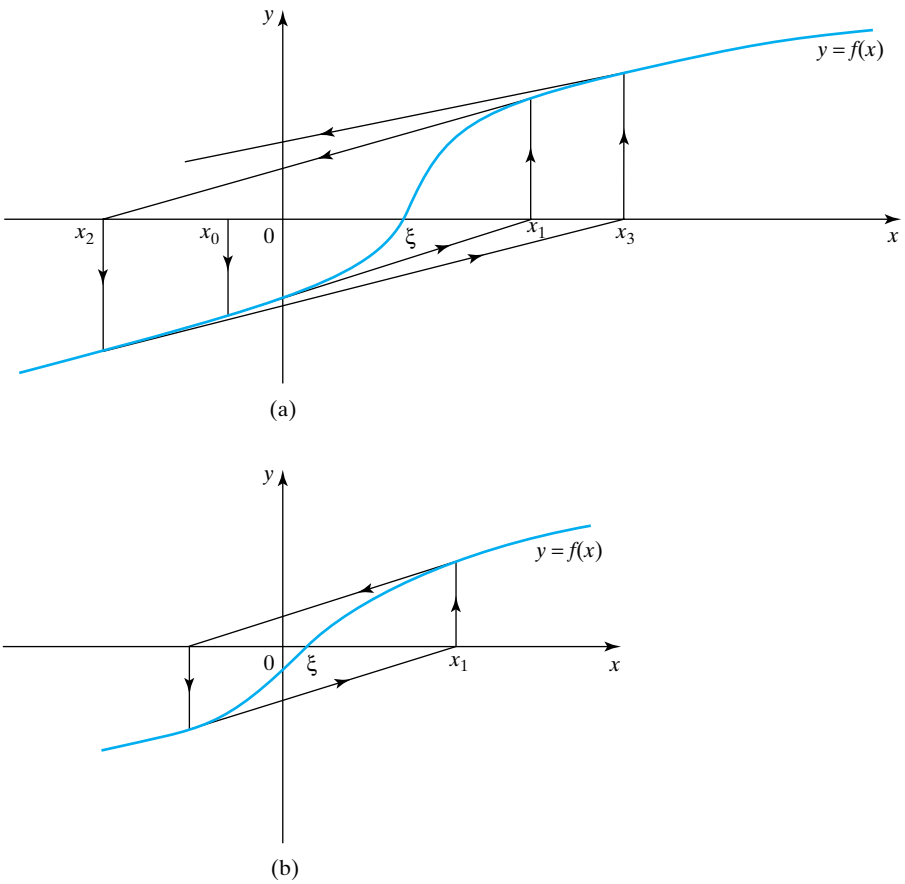


FIGURE 19.5 (a) Divergent process. (b) Repeated cycle.

ISAAC NEWTON (1642–1727)
An English mathematician and scientist who was born on Christmas day to a farming family, his father having died before he was born. His abilities as a child led to him study at Cambridge University where he later held the Lucasian Chair of mathematics. He created the forerunner of modern differential calculus, then called the theory of fluxions, by the age of 23. After a two-year stay at home to avoid a severe outbreak of the bubonic plague elsewhere in England, he returned to Cambridge in 1667 where for two years he pursued his interest in optics. The Lucasian Professorship of Mathematics was held by Barrow, who resigned it in 1669 so that Newton could be appointed. It was after this that many of his most important results were published, including his world famous *Philosophiae naturalis principia mathematica* in 1687, though many of his results were obtained long before they first appeared in print. He made contributions throughout mathematics and science and is universally recognized as one of the greatest mathematicians of all time.

Summary

The need for the determination of roots of nonlinear functions arises in many ways. The methods for the determination of roots discussed in this section were the bisection method, fixed point iteration methods, and Newton's method, which can be considered as a special fixed point iteration method. It was stressed that the bisection method only works for functions that change sign across a root, that its rate of convergence to a root is slow, and that if more than one root occurs in an interval it is not known in advance to which one the method will converge. The relative speeds of convergence of these methods were mentioned.

EXERCISES 19.2

In Exercises 1 through 6 use the bisection method to find the required root.

1. The root of $\sin x - \frac{1}{3}x = 0$ close to $x = 2.2$.
2. The root of $e^{x/3} - x^2 = 0$ close to $x = 1.1$.
3. The root of $3\ln x + x^2 - 3 = 0$ close to $x = 1.3$.
4. The largest positive root of $x^3 - 1.9x^2 - 2.3x + 3.7 = 0$.
5. The smallest root of $x^3 - 4.5x^2 + 1.3x + 8 = 0$.
6. The root of $\frac{1}{2}\sqrt{1 - x^2} - x^2 = 0$.

In Exercises 7 through 12 use a fixed point iteration scheme to find the required roots.

7. Determine $a^{1/n}$ where $a > 0$ and n is an integer. Check the result by finding $4^{1/3}$.
8. Find the roots of $x^2 + 4x + 1 = 0$ and check the results by using the quadratic formula.
9. Find all three roots of $x^3 - 4.3x^2 + 1.4x + 7.8 = 0$.
10. Find the positive root of $\sin x - \frac{1}{2}x = 0$.
11. Find the positive root of $x^2 - 2\sinh x + 1 = 0$.
12. Find the positive root of $x^2 + 2\ln x - 4 = 0$.

In Exercises 13 through 18 use Newton's method to find the required root.

13. Find $23^{1/3}$ by solving for the zero of $f(x) = 23 - x^3$.
14. Find the smallest positive root of $\tan x + 2\tanh x = 0$.
15. Find the largest root of $x^4 - 4x^3 + x^2 + 1.2 = 0$.
16. Find the smallest root of $x^4 - 3x^3 + 2x^2 - 3x - 1.6 = 0$.
17. Find the root of $3x - e^{-x} = 0$.
18. Find the root of $1 + \tanh x - 2\tan x = 0$.

19.3 Interpolation and Extrapolation

Sometimes a function $f(x)$ that is assumed to be smooth is only known in the form of a set of discrete values $y_i = f(x_i)$ at a set of arguments $x_1, x_2, \ldots, x_n$ such that $x_1 < x_1 < \cdots < x_n$. When this occurs it often becomes necessary to estimate the value $f(\alpha)$ when α lies between two of the known arguments x_i. This process is called the **interpolation** of the function $f(x)$ between its known values, and the interpolated value $f(\alpha)$ is estimated using some or all of the known values y_i. Various methods are available for interpolation, but nothing can be said about the

error involved unless some assumptions are made about the function. As a general rule the error is best reduced by selecting a method that reflects the apparent variation of $f(x)$. Some of the factors to be taken into account when choosing an interpolation method are whether $f(x)$ appears to be convex or concave for $x_1 < x < x_n$, whether it is oscillatory, and whether it exhibits sharp curvature at a point or points belonging to the interval.

The estimation of $f(\alpha)$ when α lies outside the interval, either to the left of x_1 or to the right of x_n, is called **extrapolation** of the function $f(x)$, and as the process can be liable to considerable error it should be used with care. As with interpolation, nothing can be said about errors produced by extrapolation unless some general properties of the function involved either are known or are assumed. The use of extrapolation is more frequent than might be expected. It is, for example, used in Newton's method when the curve at a point is replaced by its tangent that is then extended (extrapolated) until it intersects the x-axis, again in the numerical solution of ordinary differential equations to be discussed later, and elsewhere.

Linear Interpolation

Let the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ belonging to an unknown smooth function $y = f(x)$ be plotted on a graph. Then the simplest way to estimate the value of $y(x)$ when x lies in the interval $x_i < x < x_{i+1}$ is to join the points (x_i, y_i) and (x_{i+1}, y_{i+1}) by a straight line segment, and then to use the point on the line segment with argument x as the approximation to $y(x)$. This process is called **linear interpolation**, and it is illustrated in Fig. 19.6, where A is the point (x_i, y_i), B is the point (x_{i+1}, y_{i+1}), and the straight line segment AB has the equation $y = \tilde{y}(x)$. Then, in linear interpolation, point P on the line segment AB is used as the approximation to Q on the curve $y = f(x)$.

graphical interpretation of linear interpolation

A simple calculation shows that the straight line segment $y = \tilde{y}(x)$ representing the **linear interpolation function** between the two points (x_i, y_i) and (x_{i+1}, y_{i+1}) is given by

$$\tilde{y}(x) = \left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right)(x - x_i) + y_i, \quad \text{for } x_i < x < x_{i+1}. \tag{7}$$

If x is chosen so that either $x < x_1$ or $x > x_n$, result (7) becomes a **linear extrapolation formula** for $y = f(x)$ outside the interval $x_1 < x < x_n$.

linear extrapolation

Result (7) is useful for interpolation when the variation of x_i and y_i between adjacent data points is small, but as the formula introduces an error due to its failure

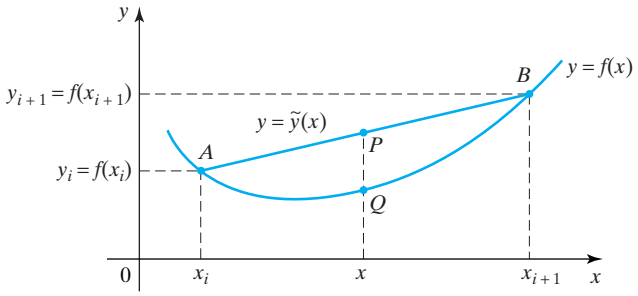


FIGURE 19.6 Linear interpolation.

to take account of the curvature of the curve, the error can become large when the result is used for extrapolation.

Lagrange Interpolation

Instead of using linear interpolation to join successive pairs of data points (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$, it is possible that a better result can be obtained by constructing a polynomial $y = P(x)$ that passes through each data point. As a polynomial is a smooth curve, it is to be hoped that it will take some account of the curvature of the function to which the data points belong, as exhibited by a set of data points, and so provide a better interpolation formula.

In Lagrange interpolation the polynomial $P(x)$ that is used is taken to be the one with the lowest possible degree that passes through each of the data points, so that when there are n data points the polynomial will be at most of degree $n - 1$. The polynomial is unique, because n equations for its n coefficients can be found by requiring it to pass through each of the n data points. The graph of this polynomial over the interval $x_1 \leq x \leq x_n$ is then used as an approximation to the unknown function $y = f(x)$ from which the data points are presumed to have been derived, on the assumption that $y = f(x)$ does not exhibit large variations as its argument x moves between the successive arguments $x_1, x_2, \ldots, x_n$ of the data points.

The polynomial $y = P(x)$ given by

$$P(x) = \sum_{k=1}^{n} L_k(x) y_k,$$

where

$$L_k(x) = \frac{(x - x_1)(x - x_2) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_1)(x_k - x_2) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}, \tag{8}$$

has the property we require, because it is of degree at most $n - 1$, and it passes through each data point, so it defines an interpolation formula over the interval $x_1 \leq x \leq x_n$. The polynomials $L_k(x)$, called **fundamental Lagrangian interpolation polynomials**, are all of degree $n - 1$, but the linear combination forming the function $P(x)$ involving the set of data points can have a lower degree. That the $L_k(x)$ have the required property is easily seen from the fact that when $x = x_k$ each $L_r(x_k)$ with $r \neq k$ contains a zero factor in its numerator so that $L_r(x_k) = 0$, but when $r = k$ we have $L_k(x_k) = 1$, showing that $P(x_k) = y_k$. The polynomial $P(x)$ provides the required **Lagrange interpolation formula** for the set of n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

fundamental Lagrangian interpolation polynomials

When $n = 2$ result (8) reduces to linear interpolation, and when $n = 3$ it becomes a quadratic, and so fits a parabola through the three points. A parabola is a smooth curve with a steadily changing gradient, so as it takes some account of the curvature of the unknown function $y = f(x)$ over the three points that are involved, it can be expected to provide a better approximation than simple linear interpolation.

However, it is inadvisable to use Lagrange interpolation over many more than three points, because when a polynomial of degree $(n - 1) \gg 1$ is forced to pass

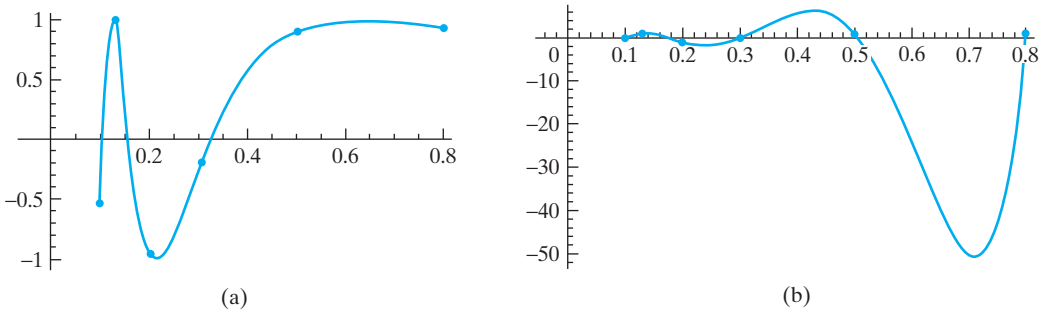


FIGURE 19.7 The function $y = f(x)$ and its Lagrange interpolation approximation $y = P(x)$ using six points.

through a set of n fixed points it usually produces a polynomial that introduces large oscillations between adjacent pairs of data points, even though the points themselves indicate no such behavior of the original function.

This undesirable characteristic of high degree Lagrange interpolation polynomials can be illustrated by constructing a fifth degree interpolation polynomial for the function

$$y(x) = \sin(1/x),$$

in the interval $0.1 \le x \le 0.8$ shown in Fig. 19.7a. When constructing an interpolation function, the precise extrema of the function are seldom known, so to reflect this uncertainty the six data points used will be the two end points and four internal points, two of which are close to, though not at, the extrema of $y(x) = \sin(1/x)$ in the interval $0.1 \le x \le 0.8$. These six data points are shown as dots on the graph of $y(x)$, and they have the following (x, y)-coordinates:

$$(0.1, -0.544021), (0.13, 0.986959), (0.2, -0.958924), (0.3, -0.190568),$$
$$(0.5, 0.909297), (0.8, 0.948985).$$

The Lagrangian interpolation polynomial that passes through these six points is

$$p(x) = -47.953442 + 1039.947347x - 7963.493901x^2 + 26828.578780x^3$$
$$-39901.683910x^4 + 21121.453960x^5.$$

The extreme oscillations that occur between the interpolation data points can be seen by inspection of Fig. 19.7b that shows the graph of $P(x)$ in the interval $0.1 \le x \le 0.8$, on which the data points are marked as dots.

In this case, as only six data points are involved, it would have been better to use three consecutive three point Lagrangian interpolation polynomials over the intervals $0.1 \le x \le 0.2$, $0.2 \le x \le 0.5$, and $0.3 \le x \le 0.8$, with the last interpolation polynomial used only in the interval $0.5 \le x \le 0.8$. However, although such a composite interpolation scheme would provide a *continuous* approximation to $y(x) = \sin(1/x)$ over the entire interval, the curve would not be smooth because of discontinuities in its derivative at $x = 0.2$ and $x = 0.5$ where the parabolic approximations meet.

We conclude this brief introduction to Lagrange interpolation by mentioning that its main use is of a theoretical nature in connection with the derivation of effective numerical techniques of various kinds. The only one of which to be developed

here is in connection with *cubic spline interpolation*, which can be considered to be a refinement of the fitting of a polynomial of low degree over two points.

Cubic Spline Interpolation

An important use of an interpolation function arises in engineering design, and elsewhere, when it becomes necessary to generate a smooth curve with an unknown equation that passes through a set of data points, without the introduction of oscillations between these points. The approach to be outlined is motivated by the old engineering drafting technique that produced such a curve by tracing along a thin flexible metal strip, called a *spline*, that by the application of pressure at points along its length was constrained to pass through each data point.

Clearly a Lagrange interpolation polynomial is unsuitable because of the oscillations it can introduce, and because in practice there may be many data points. The approach we will use instead will be to approximate the curve in a piecewise manner by a polynomial of degree 3 over each interval $x_i \le x \le x_{i+1}$ in such a way that both the first and second derivatives of the curve at the ends of the interval match those of the approximations to the immediate left at x_i and those of the approximation to the immediate right at x_{i+1}. Composite approximations of this type are called **cubic spline function** approximations. In the mathematical approach to the determination of the spline function approximation through the n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the x_i are called the **nodes** of the approximation, and the corresponding points y_i where adjacent curves meet are called the **knots** of the approximation.

cubic splines, nodes, and knots

The mathematical requirements to be satisfied by a spline function approximation are seen to be:

- **(a)** Each curve through the adjacent points (x_i, y_i) and (x_{i+1}, y_{i+1}) is a cubic.
- **(b)** The composite curve over the entire interval must interpolate the data by passing through each knot.
- **(c)** The curve itself and the first and second derivatives of the composite curve must be continuous at the nodes x_i.
- **(d)** Conditions must be prescribed at the end points x_1 and x_n of the interval, depending on whether the data points indicate that *beyond* these points the extrapolation curve is required to approach a straight line or a parabola, or to exhibit some other behavior such as periodicity over the interval $x_1 \le x \le x_n$.

Because of conditions (i) to (iii) the second derivative $f''(x)$ must vary linearly over each interval $x_i \le x \le x_{i+1}$ and be continuous across each node, so using the Lagrange interpolation formula we can write

$$f''(x) = \left(\frac{x_{i+1} - x}{x_{i+1} - x_i}\right) f''(x_i) + \left(\frac{x - x_i}{x_{i+1} - x_i}\right) f''(x_{i+1}) \quad \text{for } x_i \le x \le x_{i+1}. \tag{9}$$

Integrating this result twice with respect to x gives

$$f(x) = \frac{1}{6}\left(\frac{3x_{i+1}x^2 - x^3}{x_{i+1} - x_i}\right) f''(x_i) + \frac{1}{6}\left(\frac{x^3 - 3x_i x^2}{x_{i+1} - x_i}\right) f''(x_{i+1}) + ax + b, \quad \text{for } x_i \le x \le x_{i+1}, \tag{10}$$

where a and b are arbitrary constants of integration. As $f(x)$ is required to pass through the points (x_i, y_i) and (x_{i+1}, y_{i+1}), substituting these two conditions into

(10) determines a and b, and after setting $d_i = x_{i+1} - x_i$ we find that

$$\begin{aligned} f(x) = {} & \frac{1}{6d_i}\left[(x_{i+1}-x)^3 f''(x_i) + (x-x_i)^3 f''(x_i)\right] \\ & + \frac{1}{6d_i}\left[6y_i - d_i^2 f''(x_i)\right](x_{i+1}-x) \\ & + \frac{1}{6d_i}\left[6y_{i+1} - d_i^2 f''(x_{i+1})\right](x-x_i), \quad \text{for } x_i \le x \le x_{i+1}. \end{aligned} \tag{11}$$

To proceed further we must now find conditions determining the derivatives $f''(x_i)$ and $f''(x_{i+1})$, and this can be accomplished by using the as yet unused condition that the first derivative $f'(x)$ must be continuous across each node. To apply this condition we differentiate (11) once with respect to x, and require the derivative when $x = x_{i+1}$ in the ith interval, that is, at its *right-hand* end point, to equal the derivative when $x = x_{i+1}$ in the $(i+1)$th interval, corresponding to its *left-hand* end point, as a result of which we find that

$$d_{i-1} f''(x_{i-1}) + 2(d_{i-1} + d_i) f''(x_i) + d_i f''(x_{i+1}) = Y_i, \tag{12}$$

where

$$Y_i = 6\left(\frac{y_{i+1} - y_i}{d_i} - \frac{y_i - y_{i-1}}{d_{i-1}}\right). \tag{13}$$

Result (12) is a set of $n-2$ linear simultaneous equations for the n derivatives $f''(x_i)$, and when these are known the spline function approximation formed by the set of functions in (11) defined over the consecutive intervals $x_i \le x \le x_{i+1}$ with $i = 1, 2, \ldots, n-1$ can be constructed. It is crucial to the practical use of splines that this linear system of equations be nonsingular, and that an extremely efficient algorithm be available for solving it.

As the values of $f''(x_1)$ and $f''(x_n)$ cannot be found from the condition that $f'(x)$ is continuous across the nodes x_1 and x_n these values must be specified as additional conditions.

The choice of values for $f''(x_1)$ and $f''(x_n)$ prescribed as end conditions must be made intuitively, based on the way the data points indicate that the interpolated curve is most likely to behave (be extrapolated) *beyond* the end points of the interval $x_1 \le x \le x_n$. Three typical choices are the **natural** or **linear spline** end condition, the **parabolic spline** end condition, and **periodic spline** end conditions.

Natural or linear spline end conditions

spline end conditions

This choice of end conditions involves setting

$$f''(x_1) = f''(x_n) = 0. \tag{14}$$

These conditions are also called the *linear spline end conditions* because although the polynomial used over the intervals is a cubic, the vanishing of the second derivative at $x = x_1$ and $x = x_n$ causes the approximation to become linear beyond the ends of the interval.

Parabolic spline end conditions

This choice of end conditions involves setting

$$f''(x_1) = f''(x_2) \quad \text{and} \quad f''(x_{n-1}) = f''(x_n). \tag{15}$$

These conditions are called the *parabolic spline end conditions* because the consequence of this choice is that $f''(x)$ is constant in each of the end intervals, causing the cubic interpolation formula to reduce to a quadratic or *parabolic* approximation.

Periodic spline end conditions

If there is reason to believe that the data is periodic over the interval $x_1 \le x \le x_n$, then the following are the appropriate end conditions

$$f(x_1) = f(x_{n-1}) \quad \text{and} \quad f'(x_n) = f'(x_2). \tag{16}$$

Other end conditions can be used and, of course, a linear spline end condition may be applied at one end of an interval and a parabolic spline end condition at the other if this is appropriate. An end condition that is more important than the parabolic end condition is the condition that leads to the complete cubic spline, namely the spline that interpolates $f'(x)$ as well as $f(x)$ at both x_1 and x_n. This spline has a higher rate of convergence as the maximum step size tends to zero, and it is often implemented using a local approximation to the derivatives that preserves the higher rate of convergence.

an example of a spline approximation

The function $y(x) = \sin(1/x)$ is shown in Fig. 19.8 as the dashed curve, on which is superimposed the cubic spline approximation with natural end conditions. The six interpolation data points are shown as dots.

For more information about topics in this section see references [2.14] through [2.20].

Summary

Linear and Lagrange interpolation were defined, and the desirability of using low degree Lagrange interpolation in order to avoid the introduction of excessive oscillations between interpolation data points was illustrated by example. Extrapolation was then defined, and its attendant dangers were stressed unless something is known about the nature of the function being extrapolated. Finally, spline interpolation was introduced, which produces

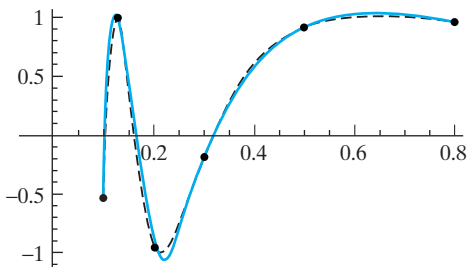


FIGURE 19.8 The function $y(x) = \sin(1/x)$, the cubic spline approximation and the data interpolation points.

a smooth interpolated curve through each data point, and the different end conditions that can be applied were explained together with their effects.

EXERCISES 19.3

Exercises in this set require the use of a computer.

1. Graph the function $f(x) = x/(1 + x^2)$ in the interval $0 \le x \le 3$. Select four points on the graph, and after constructing a polynomial that passes through each of the points graph the polynomial and compare the result with the original function.
2. Graph the function $f(x) = \sin x/(1 + x^2)$ in the interval $0 \le x \le \pi$. Select four points on the graph, and after constructing a polynomial that passes through each of the points graph the polynomial and compare the result with the original function.
3. Graph the function $f(x) = 1 + x \sin x$ in the interval $0 \le x \le 2\pi$. Select seven points on the graph, and after constructing a polynomial that passes through each of the points graph the polynomial and compare the result with the original function. Try to improve the approximation by choosing the seven points differently.
4. Graph the function $f(x) = (1 - x^5)^{1/5}$ in the interval $0 \le x \le 1$. Select seven points on the graph, and after constructing a polynomial that passes through each of the points graph the polynomial and compare the result with the original function. Try to improve the approximation by choosing the seven points differently.
5. Graph the function $f(x) = 1 - 2x \cos x$ in the interval $0 \le x \le 2\pi$. Select seven points on the graph and construct a spline function approximation to the function in the interval $0 \le x \le 2\pi$ using parabolic spline function end conditions. Graph the spline function and compare it with the graph of the original function. Repeat the calculation using linear spline function end conditions and compare the result with the previous graph.
6. Graph the function $f(x) = (1 - x^7)^{1/7}$ in the interval $0 \le x \le 1$. Select seven points on the graph, and construct a spline function approximation to the function in the interval $0 \le x \le 1$ using linear spline function end conditions. Graph the spline function and compare the result with the original function. Repeat the calculation using parabolic spline function end conditions and compare the result with the previous graph.

19.4 Numerical Integration

The need for **numerical integration**, also called **numerical quadrature**, arises when either a definite integral that is required cannot be evaluated analytically, or when special functions involved in an analytical solution are too complicated to be of direct use. A typical definite integral that can only be evaluated numerically is

$$I = \int_0^5 \frac{\sin 3x}{\sqrt{x^2 + x + 1}} dx,$$

the value of which can be shown to be $I = 0.364\,873$. In what follows, three different numerical integration schemes for the evaluation of definite integrals will be derived called, respectively, the *trapezoidal rule*, *Simpson's rule*, and *Gaussian integration*. Of these three methods the first is the least accurate, whereas the last provides high accuracy with far fewer computational steps than the frequently used Simpson's rule.

The Trapezoidal Rule

The basis of this very simple rule can be understood from Fig. 19.9 in which the integral $I = \int_a^b f(x)dx$ is approximated by the area of the *trapezoid PQRS* shown as the shaded area in the interval $a \le x \le b$ associated with the graph of $y = f(x)$.

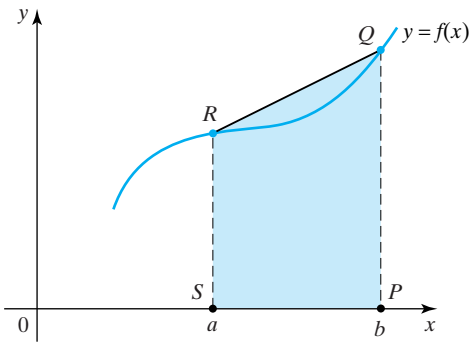


FIGURE 19.9 A trapezoidal approximation to $I = \int_a^b f(x)dx$.

As the area $PQRS = \frac{1}{2}(b-a)[f(a)+f(b)]$, the approximation to the definite integral in Fig. 19.9 is given by

$$\int_a^b f(x)dx \approx \frac{1}{2}(b-a)[f(a)+f(b)]. \tag{17}$$

Setting $b-a=h$, and denoting by $E(h)$ the error made when approximating the definite integral by a single trapezoid with base h, we have

$$E(h) = \frac{1}{2}(b-a)[f(a)+f(b)] - \int_a^b f(x)dx,$$

so in terms of $E(h)$ the approximation (17) can be replaced by the exact result

$$\int_a^b f(x)dx = \frac{1}{2}(b-a)[f(a)+f(b)] - E(h). \tag{18}$$

A different way of deriving result (17) is to use linear interpolation to represent $y(x)$ between $x=a$ and $x=b$, and then to integrate the result.

Although the exact error $E(h)$ is not known, an expression for the error can be derived on the assumption that $f(x)$ is suitably differentiable in the range of integration $a \le x \le b$. The error term for the trapezoidal rule will be stated without proof because its derivation is similar to that for the more accurate Simpson's rule, and this will be given later. When a definite integral is approximated by a single trapezoid, as in Fig. 19.9, the error term in (18) is given by $E(h) = \frac{1}{12}h^3 f''(\xi)$, for some ξ such that $a \le \xi \le b$. If we use this error term (18) becomes

$$\int_a^b f(x)dx = \frac{1}{2}(b-a)[f(a)+f(b)] - \frac{1}{12}h^3 f''(\xi), \tag{19}$$

for some ξ such that $a \le \xi \le b$.

A better estimate of the definite integral $\int_a^b f(x)dx$ can by obtained by dividing $a \le x \le b$ into n subintervals, applying (19) to each of the n subintervals, and then summing the result. Although not necessary, it is usual to choose all n subintervals to be of equal length $h = (b-a)/n$, where h is usually called the **step size**. Consequently, setting $x_i = a + ih$ for $i = 0, 1, \ldots, n$, we arrive at what is called

composite trapezoidal rule with error term

the **composite trapezoidal rule**

$$\int_a^b f(x)dx = \frac{1}{2}h\left[f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right] - \frac{1}{12}(b-a)h^2 f''(\eta), \tag{20}$$

where the unknown number η in the error term is such that $a \le \eta \le b$.

The error term in the composite trapezoidal rule is obtained from the error term in (19) by addition of the error terms in each subinterval. The details of the derivation will be left as an exercise, because they parallel those for the corresponding case in the composite Simpson's rule that will shortly be discussed in detail.

Although η is not known, whenever it is possible to estimate the greatest and least values of $f''(x)$ in the interval $a \le x \le b$, bounds can be placed on the composite trapezoidal rule result by assigning these values of $f''(x)$ to $f''(\eta)$.

In practical applications of the composite trapezoidal rule the error term is usually only used to show that as the number n of subintervals increases, so the error decreases as $(b-a)h^2/12$, where $h = (b-a)/n$. The error is often approximated by forming two approximations with different h and using the asymptotic behavior to estimate the error of the result corresponding to the smaller h. Another approach is to compare the result with the one obtained with Simpson's method.

EXAMPLE 19.5

Use the composite trapezoidal rule with $n = 10$, 30, and 50 subintervals to evaluate

$$I = \int_0^5 \frac{\sin 3x}{\sqrt{x^2 + x + 1}}dx,$$

and approximate the error when 50 subintervals are used.

Solution The following results were obtained by computer:

n	10	30	50
$I_{\text{trap}(n)}$	0.290422	0.356897	0.362010

The result for $I_{\text{trap}(50)}$ should be compared with the result $I = 0.364873$ obtained by a higher order method that is known to be correct to six decimal places.

Instead of using $f''(\eta)$ when approximating the error with $n = 50$, where η is unknown, we will use the easily computed average f''_{av} of $f''(x)$ over the interval, where

$$f''_{av} = \frac{1}{b-a}\int_a^b f''(x)dx = \frac{1}{b-a}[f'(b) - f'(a)].$$

We have $b - a = 5$ and the step size $h = 5/50 = 0.1$, so

$$f''_{av} = \frac{1}{5}\int_0^5 f''(x)dx = \frac{1}{5}[f'(5) - f'(0)] = -0.686.$$

Using f''_{av} in the error term instead of $f''(\eta)$ leads to $\frac{1}{12} \cdot 5 \cdot (0.1)^2 \cdot (-0.686) = -0.002858$ as the approximation to the error. Consequently, allowing for this error, the estimate of the integral is $0.362010 - (-0.002858) = 0.364868$. When this is compared with the result $I = 0.364873$ we see that in this case the error approximation is good. ■

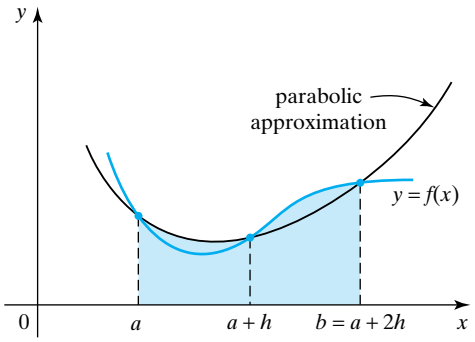


FIGURE 19.10 A parabolic approximation to $I = \int_a^b f(x)dx$.

Simpson's Rule

In its simplest form, the trapezoidal rule applied to $I = \int_a^b f(x)dx$ represents $f(x)$ by the single trapezoidal area $PQRS$ shown in Fig. 19.9, where in the interval $a \le x \le b$ the function $y = f(x)$ is approximated by the straight line segment QR. A more accurate result would be expected if a point on the curve $y = f(x)$ is chosen inside the interval $a \le x \le b$ and $f(x)$ is approximated by a parabola that passes through the two end points and the single internal point, as shown in Fig. 19.10.

Setting $b = a + 2h$, where h is the step size, and taking the additional point in the interval of integration to be $x = a + h$, so it is *midway* between the ends of the interval, the parabola to be fitted must pass through the three consecutive points $(a,\ f(a))$, $(a + h,\ f(a + h))$, and $(a + 2h,\ f(a + 2h))$. The Lagrange interpolation formula that fits a quadratic through these three points is

$$L(x) = \frac{1}{2}\frac{(x - a - h)(x - a - 2h)}{h^2} f(a) - \frac{(x - a)(x - a - 2h)}{h^2} f(a + h) + \frac{1}{2}\frac{(x - a)(x - a - h)}{h^2} f(a + 2h). \tag{21}$$

Integrating $L(x)$ over the interval $a \le x \le a + 2h$ and simplifying the result gives

$$\int_a^{a+2h} f(x)dx \approx \frac{1}{3}h[f(a) + 4f(a + h) + f(a + 2h)], \tag{22}$$

which is the result known as **Simpson's rule**, or sometimes **Simpson's 1/3 rule**. Result (22) can also be written in terms of the limits of integration a and $b = a + 2h$ as

$$\int_a^{a+2h} f(x)dx \approx \frac{1}{6}(b - a)\left[f(a) + 4f\left(\frac{a + h}{2}\right) + f(b)\right]. \tag{23}$$

If the error in Simpson's rule is denoted by $E(h)$, the approximation in (22) can be replaced by the exact result

$$\int_a^{a+2h} f(x)dx = \frac{1}{3}h[f(a) + 4f(a+h) + f(a+2h)] - E(h). \tag{24}$$

We will now derive an expression for $E(h)$ but before doing so, in order to simplify the manipulation, it will be convenient to write the limits of integration in the more symmetrical form $a = c - h$ and $b = c + h$. In terms of c and h (24) becomes

$$E(h) = \frac{1}{3}h[f(c-h) + 4f(c) + f(c+h)] - \int_{c-h}^{c+h} f(x)dx.$$

We now differentiate this result with respect to h to obtain

$$\begin{aligned} E'(h) = {} & \frac{1}{3}[f(c-h) + 4f(c) + f(c+h)] \\ & + \frac{1}{3}h[-f'(c-h) + f'(c+h)] - [f(c+h) + f(c-h)], \end{aligned}$$

where the last group of terms on the right follow from differentiating the definite integral using Leibniz's theorem (Theorem 1.5). If we set $h = 0$, this result shows that $E'(0) = 0$.

Differentiation of $E'(h)$ gives

$$E''(h) = \frac{1}{3}[f'(c-h) - f'(c+h)] + \frac{1}{3}h[f''(c+h) + f''(c-h)],$$

so setting $h = 0$ we find that $E''(0) = 0$. One final differentiation gives

$$E'''(h) = \frac{1}{3}h[f'''(c+h) - f'''(c-h)],$$

but this can be simplified by using the Taylor expansion of $f'''(c+h)$ with a remainder after the first term, where the expansion is about the point $c - h$, so that

$$f'''(c+h) = f'''(c-h) + 2hf^{(4)}(\xi),$$

where ξ is unknown but lies in the interval $c - h < \xi < c + h$.

The error term can now be found by integrating this last result three times using the results $E'(0) = E''(0) = 0$.

We have

$$\int_0^h E'''(t)dt = E''(h) - E''(0) = E''(h),$$

so

$$E''(h) = \frac{2}{3}f^{(4)}(\xi)\int_0^h t^2 dt = \frac{2}{9}h^3 f^{(4)}(\xi).$$

A further integration using the result

$$\int_0^h E''(t)dt = E'(h) - E'(0) = E'(h)$$

gives

$$E'(h) = \frac{2}{9} f^{(4)}(\xi) \int_0^h t^3 dt = \frac{1}{18} h^4 f^{(4)}(\xi).$$

Finally, after another integration we arrive at the result

$$E(h) = \frac{1}{18} f^{(4)}(\xi) \int_0^h t^4 dt = \frac{1}{90} h^5 f^{(4)}(\xi), \tag{25}$$

which is the required expression for the error term. Using this result in (24) gives

$$\int_a^{a+2h} f(x)dx = \frac{1}{3}h[f(a) + 4f(a+h) + f(a+2h)] - \frac{1}{90}h^5 f^{(4)}(\xi). \tag{26}$$

As $f^{(4)}(\xi)$ enters as a factor in $E(h)$, this shows the rather surprising result that although Simpson's rule was derived by requiring a quadratic polynomial to pass through three points, the rule is actually exact for cubic polynomials.

As with the trapezoidal rule, the accuracy of Simpson's rule can be improved by increasing the number of subintervals, but as the rule is equivalent to constructing parabolas through three consecutive equispaced points, to use the rule over more than three points the number of points chosen for the interval $a \le x \le b$ must be *odd*, so that the number of intervals must be *even*. Dividing the interval $a \le x \le b$ into $2n$ equal subintervals each of length $h = (b-a)/2n$, and adding the results, gives the **composite Simpson's rule**

composite Simpson's rule with error term

$$\int_a^b f(x)dx = \frac{1}{3}h\left[f(a) + 4\sum_{i=1}^{n} f(a+(2i-1)h) + 2\sum_{i=1}^{n-1} f(a+2ih) + f(b)\right] - \frac{1}{180}(b-a)h^4 f^{(4)}(\eta) \tag{27}$$

where η is unknown but is such that $a < \eta < b$.

The error term in the composite rule (27) is obtained as follows. Let $x_i = a + 2ih$, with $i = 0, 1, \ldots, n$, and let ξ_i be the value of ξ in the interval $x_i \le x \le x_{i+1}$ appropriate to the Simpson's rule applied to that interval. Consequently, when the composite Simpson's rule is formed, the error term in each of these intervals will be added. Now, each derivative $f^{(4)}(\xi_i)$ must satisfy the inequality

$$\min_{a \le x \le b} f^{(4)}(x) \le f^{(4)}(\xi_i) \le \max_{a \le x \le b} f^{(4)}(x),$$

so the addition of these n results followed by division by n gives

$$\min_{a \le x \le b} f^{(4)}(x) \le \frac{1}{n}\sum_{i=1}^{n} f^{(4)}(\xi_i) \le \max_{a \le x \le b} f^{(4)}(x).$$

error estimation for composite Simpson's rule

Finally, assuming $f^{(4)}(x)$ is continuous, it follows from the intermediate value theorem that some number η exists, with $a < \eta < b$, such that

$$f^{(4)}(\eta) = \frac{1}{n}\sum_{i=1}^{n} f^{(4)}(\xi_i).$$

If we use the result $h = (b-a)/2n$, the error term in the composite Simpson's rule is seen to be given by

$$-\frac{1}{90}h^5\sum_{i=1}^{n} f^{(4)}(\xi_i) = -\frac{1}{180}(b-a)h^4 f^{(4)}(\eta).$$

EXAMPLE 19.6

Use the composite Simpson's rule with $n = 10$, 30, and 50 subintervals to evaluate

$$I = \int_0^5 \frac{\sin 3x}{\sqrt{x^2+x+1}}dx$$

and compare the results obtained with the result $I = 0.364873$, which is accurate to six decimal places. Compare the result of integrating this definite integral by the trapezoidal rule and Simpson's rule.

Solution The following results were obtained by computer:

n	10	30	50
$I_{\text{simp}(n)}$	0.376738	0.365019	0.364892

Comparison of the result $I = 0.364873$, known to be correct to six decimal places, with $I_{\text{simp}(50)} = 0.364892$, shows that $I_{\text{simp}(50)}$ only overestimates the true result by 0.000025.

When comparing the composite Simpson's rule with the composite trapezoidal rule it should be remembered that Simpson's rule subdivides the interval of integration into $2n$ subintervals, whereas the composite trapezoidal rule only uses n subintervals. The following computer results provide a comparison on this basis:

n	20	40	60	80	100
$I_{\text{trap}(n)}$	0.346825	0.360395	0.362886	0.363756	0.364158
$I_{\text{simp}(n/2)}$	0.376738	0.365626	0.365019	0.364919	0.364892

Gaussian Quadrature

Many more numerical integration methods exist than have been outlined so far, but the only other important one to be mentioned here is due to C. F. Gauss. He showed that if, when evaluating numerically an integral in the standard form

$$\int_{-1}^{1} f(x)dx,$$

the points x_i at which the values of the integrand $f(x)$ are sampled are chosen in a special way, then when n sample points are used the result can be made exact in the case that $f(x)$ is an arbitrary polynomial of degree $2n-1$ or less. Unlike Simpson's

rule, in this method the n sample points x_i are *nonuniformly* spaced throughout the interval of integration $-1 \le x \le 1$, and they are all contained *inside* the interval.

The sample points, or **nodes** as they are called, are chosen to get a formula that will integrate exactly polynomials of as high degree as possible. It turns out that the n sample points are real and lie in the open interval $(-1, 1)$, and polynomials of degree $2n - 1$ are integrated exactly.

A somewhat different approach to integration involves specifying some of the sample points to be used, and then trying to find the remaining ones so as to integrate polynomials of as high degree as possible. Formulas of this type that evaluate function values at the two ends of the interval of integration are called **Lobatto** formulas, and the trapezoidal rule and Simpson's rule are formulas of the lowest order that belong to this family.

The point is that if it is useful to specify sample points at the end points of an interval of integration it is possible to proceed in this way. However, as would be expected, if this approach is adopted it is not possible to get a formula that is as accurate as one in which no constraint is placed on the sample points.

The previous arguments are all based on the assumption that functions are approximated by (algebraic) polynomials, though sometimes it is more natural to approximate them by trigonometric polynomials (finite Fourier series).

The composite trapezoidal rule is, in fact, the optimal formula of Gaussian integration type based on trigonometric approximation. As a result it converges faster than any power of h when applied to a periodic analytic function over a multiple of a period, so for this reason it is used to compute Fourier coefficients.

To illustrate the approach used to obtain this type of integration formula, we consider the simplest situation in which $n = 2$, so as only the two sample points x_1 and x_2 are involved, with $-1 < x_1 < x_2 < 1$, the integration formula becomes

$$\int_{-1}^{1} f(x)dx \approx w_1 f(x_1) + w_2 f(x_2).$$

weights in integration formula

At this stage the values of the two sample points x_1 and x_2 are unknown, as are the numbers w_1 and w_2, called the **weights** for the integration formula at these sample points. To determine these four numbers we impose the requirement that this formula should be exact when $f(x)$ is an arbitrary polynomial of degree $2n - 1 = 3$ or less.

Let $f(x)$ be the cubic polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3,$$

in which the coefficients c_0, c_1, c_2, and c_3 are arbitrary. Then for the integration to be exact, the numbers x_1, x_2, w_1, and w_2 must be such that

$$\int_{-1}^{1} (c_0 + c_1 x + c_2 x^2 + c_3 x^3)dx = w_1 \left(c_0 + c_1 x_1 + c_2 x_1^2 + c_3 x_1^3\right) + w_2 \left(c_0 + c_1 x_2 + c_2 x_2^2 + c_3 x_2^3\right).$$

Evaluating the integral on the left, and equating the respective multipliers of the arbitrary coefficients c_0, c_1, c_2, and c_3 to make this result an identity, leads to the

results

$$\begin{aligned}
&\text{(coefficient } c_0\text{)} && w_1 + w_2 = \int_{-1}^{1} dx = 2\\
&\text{(coefficient } c_1\text{)} && w_1x_1 + w_2x_2 = \int_{-1}^{1} x\,dx = 0\\
&\text{(coefficient } c_2\text{)} && w_1x_1^2 + w_2x_2^2 = \int_{-1}^{1} x^2dx = \tfrac{2}{3}\\
&\text{(coefficient } c_3\text{)} && w_1x_1^3 + w_2x_2^3 = \int_{-1}^{1} x^3dx = 0.
\end{aligned}$$

This set of equations has the unique solution $x_1 = -1/\sqrt{3}$, $x_2 = 1/\sqrt{3}$, $w_1 = 1$, and $w_2 = 1$. Consequently, when $n = 2$, we have

The sample points

$$x_1 = -1/\sqrt{3}, \quad x_2 = 1/\sqrt{3},$$

The weights

$$w_1 = 1, \quad w_2 = 1,$$

so the extremely simple two-point integration formula that gives exact results when $f(x)$ is a polynomial of degree 3 or less is seen to be given by

$$\int_{-1}^{1} f(x)dx = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

When this approach is extended to n points, an examination of the derivation of the formula shows that the sample points $x_1, x_2, \ldots, x_n$ are simply the n roots of the Legendre polynomial $P_n(x) = 0$ of degree n, with the corresponding weight w_i at x_i given by $w_i = 2[P'(x_i)]^2/(1 - x_i^2)$, for $i = 1, 2, \ldots, n$. The general integration formula involving n points becomes

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

Gauss–Legendre integration formulas

and, collectively, these results are called **Gaussian** integration formulas or, sometimes, **Gauss–Legendre integration formulas**. It can be shown that the remainder term that must be added to the right-hand side of this last result for it to be exact for *any* function $f(x)$ with a continuous derivative $f^{(2n)}(x)$ is $R_n = \frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} f^{(2n)}(\xi)$, for some unknown ξ such that $-1 < \xi < 1$. A list of Gaussian sampling points x_i and their associated weights w_i is given in Table 19.1 for $n = 2, 3, 4, 5, 10$, and 16.

error term in Gaussian integration

As would be expected, if $f(x)$ is an arbitrary polynomial of degree $2n - 1$ or less, it follows directly that $R_n \equiv 0$, confirming that in this case the result is exact.

TABLE 19.1 Gaussian Sampling Points and Weights

n	i	x_i	w_i
2	1	−0.57735 02692	1.00000 00000
	2	0.57735 02692	1.00000 00000
3	1	−0.77459 66692	0.55555 55556
	2	0.00000 00000	0.88888 88889
	3	0.77459 66692	0.55555 55556
4	1	−0.86113 63115	0.34785 48451
	2	−0.33998 10436	0.65214 51548
	3	0.33998 10436	0.65214 51548
	4	0.86113 63115	0.34785 48451
5	1	−0.90617 98459	0.23692 68851
	2	−0.53846 93101	0.47862 86705
	3	0.00000 00000	0.56888 88889
	4	0.53846 93101	0.47862 86705
	5	0.90617 98459	0.23692 68851
10	1	−0.97390 65285	0.06667 13443
	2	−0.86506 33667	0.14945 13492
	3	−0.67940 95683	0.21908 63625
	4	−0.43339 53941	0.26926 67193
	5	−0.14887 43390	0.29552 42247
	6	0.14887 43390	0.29552 42247
	7	0.43339 53941	0.26926 67193
	8	0.67940 95683	0.21908 63625
	9	0.86506 33667	0.14945 13492
	10	0.97390 65285	0.06667 13443
16	1	−0.98940 09350	0.02715 24594
	2	−0.94457 50231	0.06225 35230
	3	−0.86563 12024	0.09515 85117
	4	−0.75540 44084	0.12462 89713
	5	−0.61787 62444	0.14959 59888
	6	−0.45801 67777	0.16915 65194
	7	−0.28160 35508	0.18260 34150
	8	−0.09501 25098	0.18945 06105
	9	0.09501 25098	0.18945 06105
	10	0.28160 35508	0.18260 34150
	11	0.45801 67777	0.16915 65194
	12	0.61787 62444	0.14959 59888
	13	0.75540 44084	0.12462 89713
	14	0.86563 12024	0.09515 85117
	15	0.94457 50231	0.06225 35239
	16	0.98940 09350	0.02715 24594

The apparent restriction of the integration to the standard interval $-1 \leq x \leq 1$ is unimportant, because if the integral involved is

$$I = \int_a^b f(x)dx,$$

where a and b are finite, the simple change of variable

$$x = \frac{1}{2}(b+a) + \frac{1}{2}(b-a)u$$

converts the integral to

$$I = \frac{b-a}{2}\int_{-1}^{1} F(u)du,$$

where $F(u)$ is the function $f(x)$ after the change of variable.

The accuracy obtained when using an n-point Gaussian integration formula depends on the extent to which the integrand can be approximated by a polynomial of degree $2n-1$. To illustrate matters, we apply the five-point formula to the following integral for which there is an analytical solution that can be used for comparison:

$$I = \int_0^{1/2} \frac{dx}{(1-x^2)^{1/2}} = \text{Arcsin}(1/2) = \frac{\pi}{6} = 0.523599.$$

The change of variable $x = \frac{1}{4}(1+u)$ maps the interval $0 \le x \le \frac{1}{2}$ onto the interval $-1 \le u \le 1$, so as $dx/du = \frac{1}{4}$, after changing the variable

$$I = \int_{-1}^{1} \frac{du}{(15-2u-u^2)^{1/2}}.$$

Setting $f(u) = 1/(15-2u-u^2)^{1/2}$ and applying the five-point Gaussian formula gives

$$I \approx 0.236927\, f(-0.906180) + 0.478629\, f(-0.538469) + 0.568889\, f(0)$$
$$+ 0.478629\, f(0.538469) + 0.236927\, f(0.906180) = 0.523599.$$

In this case the numerical approximation is seen to be correct to six decimal places.

modern adaptive integration codes

The key idea used in modern integration codes involves the use of an adaptive algorithm. In such codes the error of an integral evaluated over an interval is approximated by comparing it to the result obtained by a higher order formula. Thus, the error of the trapezoidal rule can be estimated by comparing the result to the one obtained using Simpson's rule. If the result is not sufficiently accurate, the interval is split in half and the two intervals are then treated separately. Reducing the length of an interval produces a significant reduction in the error. This can be seen by considering the low-order trapezoidal rule. The effect of halving the interval is to reduce the error in a half interval by a factor of approximately an eighth, so as the operation of integration is linear, the error over the full interval is reduced by a factor of approximately a fourth. When this argument is extended, we see that if the interval of integration is divided into many pieces, accurate values of the integral over all the pieces can be added together to get an accurate value over the whole interval, with the same being true of the error estimates.

In this approach two formulas are applied to an interval using as many values of f as possible in both formulas. That the method is computationally efficient when the combination of the trapezoidal rule is used and Simpson's rule is used can be seen from the fact that only one extra evaluation of f is necessary in order to estimate the error. Modern codes use a Gaussian formula of high order as the basic formula, and a special formula of much higher order that makes use of as many function evaluations as possible for estimating the error.

As Gaussian integration formulas make no use of the values of the integrand at the end points of the interval $-1 \le x \le 1$, they can be used to approximate a convergent improper integral of the type $\int_a^b f(x)dx$, where the integrand becomes infinite at either end point.

For more information about numerical integration, see references [2.14] through [2.20].

Summary

The methods for numerical integration, also called numerical quadrature, introduced in this section were the trapezoidal and composite trapezoidal rule, Simpson's rule and the composite Simpson's rule, and Gaussian quadrature. The relative accuracies of the methods were explained; the high accuracies of Gaussian quadrature was stressed. The suitability of the composite trapezoidal rule for the computation of Fourier coefficients was mentioned.

EXERCISES 19.4

The following exercises require the use of a computer.

1. Use the composite Simpson's rule with step length $h = 0.5$ to determine
$$I = \int_1^3 (2x^3 - 3x^2 + 4x - 1)dx,$$
and hence verify that the rule integrates cubics exactly.
2. Use the composite trapezoidal rule with the step length $h = 0.1$ to evaluate
$$I = \int_0^1 \frac{dx}{1+x^2},$$
and estimate the error term involved. Compare your results with the exact value $I = \frac{\pi}{4}$. Repeat the calculation using the composite Simpson's rule with the same step length, but without estimating the error.
3. Use the composite trapezoidal rule and Simpson's rule, each with 10 subintervals, to estimate
$$I = \int_0^\pi \frac{\sin x}{x}dx,$$
and compare your results with $I = 1.851937$, which is exact to six decimal places.
4. Use the composite trapezoidal and Simpson's rule, each with step length $h = 0.2$, to estimate
$$I = \int_0^2 x^2 e^{-x}dx,$$
and compare your results with the analytical solution $I = \frac{1}{13} + \frac{1}{2}\,\text{Arctan}\,5 - \frac{1}{8}\pi$.
5. Use the composite Simpson's rule with step length $h = 0.4$ to estimate
$$I = \int_2^6 \frac{\ln(2+3\sqrt{x})}{1+x^2}dx,$$
and compare your result with the result $I = 0.596545$, which is correct to six decimal places.
6. Use the composite trapezoidal and Simpson's rule, each with step length $h = 0.4$, to estimate
$$\int_0^4 \left(1 - \frac{x}{4}\right)^4 x^{1/2}dx.$$
Compare your results with the exact solution that follows from the general result
$$I(z,n) = \int_0^n \left(1 - \frac{x}{n}\right)^n x^{z-1}dx = \frac{1\cdot 2\cdot 3\cdots n}{z(z+1)(z+2)\cdots(z+n)}n^z.$$
It follows from the definition of the gamma function that $\lim_{n\to\infty} I(z,n) = \Gamma(z)$. Explain why replacing 4 by 50 in the original integral and evaluating the result using the composite Simpson's rule with more subdivisions is not likely to lead to much improvement of the poor estimate it provides of $\Gamma(3/2) = \frac{1}{2}\sqrt{\pi}$.
7. The Bessel function $J_1(x)$ has the integral representation
$$J_1(x) = \frac{2}{\pi}\int_0^{\pi/2} \sin(x\cos\theta)\cos\theta\, d\theta.$$
Use the composite Simpson's rule with step length $h = \pi/20$ to estimate $J_1(2)$, and compare your result with the result $J_1(2) = 0.576725$, which is accurate to six decimal places.

In Exercises 8 through 10 use the integral representation
$$J_n(x) = \frac{1}{\pi}\int_0^\pi \cos(x\sin\theta - n\theta)d\theta.$$

8. Estimate $J_2(2)$ using the composite Simpson's rule with step length $h = \pi/8$, and compare your result with $J_2(2) = 0.352834$, which is accurate to six decimal places.

9. Estimate $J_1(4)$ using the composite Simpson's rule with step length $h = \pi/10$, and compare your result with $J_1(4) = -0.066043$, which is accurate to six decimal places.

10. Estimate $J_3(4)$ using the composite Simpson's rule with step length $h = \pi/10$, and compare your result with $J_3(4) = 0.430171$, which is accurate to six decimal places.

11. The modified Bessel function $I_0(x)$ has the integral representation

$$I_0(x) = \frac{2}{\pi}\int_0^{\pi/2} \cosh(x\sin\theta)d\theta.$$

Use the composite Simpson's rule with step length $h = \pi/16$ to estimate $I_0(3.5)$, and compare your result with $I_0(3.5) = 7.378203$, which is correct to six decimal places.

12. The modified Bessel function $I_1(x)$ has the integral representation

$$I_1(x) = \frac{2x}{\pi}\int_0^{\pi/2} \cosh(x\sin\theta)(\cos\theta)^2 d\theta.$$

Use the composite Simpson's rule with step size $h = \pi/16$ to estimate $I_1(3)$, and compare your result with $I_1(3) = 3.953370$, which is correct to six decimal places.

In Exercises 13 and 16 use the 3-, 5-, and 10-point Gaussian formulas to estimate the given integral and compare the results with the exact value.

13. $I = \displaystyle\int_0^{3\pi/2} \cos x dx.$

The exact value is $I = -1$.

14. $I = \displaystyle\int_0^{\pi/2} e^{-x}\cos x.$

The exact value to six decimal places is $I = \frac{1}{2}[1 + \exp(-\frac{1}{2}\pi)] = 0.603940$.

15. Use the 10-point Gaussian formula to estimate the value of the convergent improper integral

$$I = \int_0^{1/2} \frac{dx}{(1-4x^2)^{1/2}}.$$

Compare the result with the exact value to six decimal places $I = \frac{\pi}{4} = 0.785398$.

18. Use the 10-point Gaussian formula to estimate the value of the convergent improper integral

$$I = \int_0^{\pi/2} \frac{\sqrt{x}}{\sin x} dx.$$

Compare the result with the exact value to six decimal places $I = 2.753142$.

19.5 Numerical Solution of Linear Systems of Equations

This section describes two approaches to the solution of systems of n nonhomogeneous linear equations in the n unknowns $x_1, x_2, \ldots, x_n$, both of which are important. These methods, with various refinements, are all found in major linear algebra software packages.

The first method, involving the successive elimination of unknowns, is of the *direct type*, in which the solution is obtained after systematically eliminating $n-1$ of the n unknowns to find x_n. The process of back-substitution is then used to find the remaining unknowns in the reverse order $x_{n-1}, x_{n-2}, \ldots, x_1$. This method can also be used when the number of equations is not equal to the number of the unknowns, when it also shows automatically if a system of equations is inconsistent.

A related method is essentially the same as the first, apart from the way in which details of the elimination process are recorded to permit solving conveniently more than one system of equations with the same coefficient matrix. It applies to systems in which the number of equations equals the number of unknowns. The approach is to attempt to factorize the coefficient matrix $\mathbf{A}$ in the system $\mathbf{Ax} = \mathbf{b}$ into the product $\mathbf{PA} = \mathbf{LU}$, where $\mathbf{L}$ is a *lower-triangular* matrix with 1's on its leading diagonal, $\mathbf{U}$ is an upper-triangular matrix, and $\mathbf{P}$ is a permutation matrix, the reason for which will be explained later. The method uses this factorization to

determine the solution vector $\mathbf{x}$. A failure of the method to achieve this factorization indicates that $\mathbf{A}$ is singular, so then one or more of its rows is linearly dependent on other rows.

The second type of approach is an *iterative* one, and it only applies to a system of n nonhomogeneous equations in the n unknowns $x_1, x_2, \ldots, x_n$. The methods start with an arbitrary approximation $\mathbf{x}^{(0)}$ to the solution vector $\mathbf{x}$, and this is iterated in such a way that it leads to successive improved approximations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}, \mathbf{x}^{(i+1)}$ to $\mathbf{x}$. The iterative process is terminated after N iterations as soon as the two successive iterates $\mathbf{x}^{(N-1)}$ and $\mathbf{x}^{(N)}$ yield approximations $x_i^{(N-1)}$ and $x_i^{(N)}$ to x_i, for $i = 1, 2, \ldots, n$, that differ by less than a small preassigned number $\varepsilon > 0$, called the **tolerance**. The final iterate is taken to be the solution of the system of equations to within the chosen tolerance. The number of iterations necessary to arrive at this approximation to the solution vector is indeterminate, because it depends on the structure of the equations, the iterative scheme involved, and the tolerance.

tolerance in iterations

As all methods of the direct type are, in a sense, derived from the standard **Gaussian elimination** process, it will be sufficient to describe this process in some detail. Later a comment will be offered concerning a modification that must be made to the process to ensure that the elimination procedure does not fail unnecessarily, and that round-off errors are minimized. The second direct method retains information contained in the Gaussian elimination process and uses it to derive the factorization $\mathbf{PA} = \mathbf{LU}$, after which the result is used to solve the system $\mathbf{Ax} = \mathbf{b}$. This method is useful when solutions are required to a system $\mathbf{Ax} = \mathbf{b}$ for a sequence of nonhomogeneous vectors $\mathbf{b}$ while leaving the coefficient matrix $\mathbf{A}$ unchanged. This can happen, for example, in the analysis of forces in a structure due to changes in loading, where the matrix $\mathbf{A}$ representing the structure stays the same, while the loading represented by the vector $\mathbf{b}$ is altered repeatedly.

Of the various iterative schemes that are available, we describe only the **Jacobi** and **Gauss–Seidel** schemes. These are widely used, though for somewhat different purposes, and they are applicable to systems of equations that possess a property called **diagonal dominance** that will be described later. Iterative methods are used when working with large matrices, where it frequently happens that many zero elements are present, often occurring in diagonal bands parallel to the leading diagonal of matrix $\mathbf{A}$. Matrices of this type are called **sparse matrices**, and they arise when solving partial differential equations, in spline interpolation, and in many other applications of matrices. More information about refinements to the Gaussian elimination process and about iterative methods in general can be found in the references cited at the end of the section.

The Gaussian Elimination Process

Let us assume that the system of equations to be solved is of the form

$$\mathbf{Ax} = \mathbf{b}, \tag{28}$$

where $\mathbf{A} = [a_{ij}]$ is an $n \times n$ matrix with constant coefficients, the column vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$ is the required solution vector, and the column vector $\mathbf{b} = [b_1, b_2, \ldots, b_n]^{\mathrm{T}}$ contains the constant nonhomogeneous terms, not every one of which is zero.

When written out explicitly, (28) becomes

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}. \tag{29}$$

It was shown in Chapter 3 that (29), equivalently (28), possesses a unique solution provided the rank of matrix $\mathbf{A}$ and the rank of the augmented matrix $[\mathbf{A}|\mathbf{b}]$ are both equal to n, in which case the formal solution of (28) can be written $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. However, the need to find different ways of calculating $\mathbf{x}$ arises from the fact that solutions in terms of the inverse matrix are *not* practicable when n is large, because of the magnitude of the task of calculating $\mathbf{A}^{-1}$ when n is large.

In both machine and hand computation, the foregoing full matrix form of the system in (29) is abbreviated to the augmented matrix, and the calculations are then performed on its entries. The augmented array corresponding to (29) is

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right]. \tag{30}$$

In this abbreviated notation the coefficients of $x_1, x_2, \ldots, x_n$ in each equation are identified by their position in the array, so the coefficient of x_1 in the second equation is a_{12}, while the coefficient of x_2 in the nth equation is a_{n2}.

Gaussian elimination

As individual equations can be scaled by a number k, and a multiple of an equation can be added to another equation, all without altering the solution, it follows that these same operations can be performed on the array in (30). The basic **Gaussian elimination process** makes use of these properties. The first stage of the elimination process involves assuming $a_{11} \neq 0$, multiplying the first row by a_{21}/a_{11}, and subtracting the result from the second row, when its first entry becomes zero. The next step is to multiply the first row by a_{31}/a_{11} and subtract the result from the third row, to make its first entry zero. A repetition of this process $n-1$ times completes the first stage of the process, after which all entries below a_{11} are zero, causing (30) to become

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{bmatrix}, \tag{31}$$

where the $a_{ij}^{(1)}$ and $b_i^{(1)}$ represent the modified elements a_{ij} and b_i after subtraction of the multiple of the corresponding elements in the first row.

The second stage of the elimination process involves assuming $a_{22}^{(1)} \neq 0$, subtracting suitable multiples of the modified second row in (31) from the $n-2$ rows below it to make all entries in the column below $a_{22}^{(1)}$ zero. A continuation of this